Matrix Approximation with Side Information: When Column Sampling is Enough

Jeongmin Chae†, Praneeth Narayanamurthy†, Selin Bac*, Shaama Mallikarjun Sharada* and Urbashi Mitra†

***Abstract*—A novel matrix approximation problem is considered herein: observations based on a few fully sampled columns and quasi-polynomial structural side information are exploited. The framework is motivated by quantum chemistry problems wherein full matrix computation is expensive, and partial computations only lead to column information. The proposed algorithm successfully estimates the column and row space of a true matrix given a priori structural knowledge of the true matrix. A theoretical spectral error bound is provided, which captures the possible inaccuracies of the side information. The error bound proves it scales in its signal-to-noise (SNR) ratio as SNR^{-1}. The proposed algorithm is validated via simulations which enable the characterization of the amount of information provided by the quasi-polynomial side information.**

I. Introduction

Classical matrix completion imputes missing entries of a matrix by exploiting structural information such as a low-rank structure. Initial work presumed uniform random sampling [2], [3]. Herein, we examine two key advances: non-uniform sampling and additional structural side information. We note that additional structural information such as a particular basis has been previously considered [4]–[10] for matrix completion with noisy or noise-free samples, but still assumes sampling uniformly at random. Recent work suggests that matrix completion methods are needed for applications wherein non-uniform sampling is required such as computer vision, bioinformatics and economics [11]–[13]. In particular, our motivating application of rate reaction computation in quantum chemistry [14]–[16], only allows for sampling of full columns of the true matrix as seen in Figure 1. While the full matrix can be computed at the expense of extremely high computational complexity, we use matrix completion as a technique to strongly reduce the computational complexity.

Recent efforts have considered forms of non-uniform sampling [13], [17]–[19], but still fall short of being applicable to our scenario of full-column samples only. In our proposed approach, we shall consider low-rank approximations of matrices that are, in fact, high rank. In [13], rank and coherence measures are examined and proposed approaches assume that the observed entries are concentrated around the main diagonal of the true matrix. Unfortunately, neither [17], nor [13] allow for full column sampling.

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

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

A part of this work was presented at IEEE ICASSP 2023 [1].

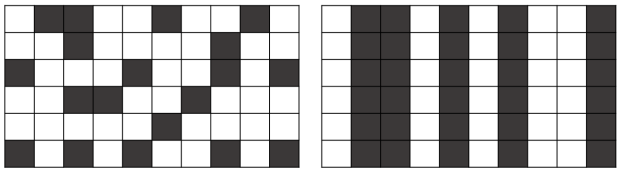

Fig. 1. An example of random sampling and column-based sampling for a matrix $\mathbf{M} \in \mathbb{R}^{6\times 10}$. The black entries indicate observed samples.

Our goal in this work is to construct a low rank-r approximation, $\hat{\mathbf{M}}$, of the true matrix $\mathbf{M}$ which has rank k where $r \leq k$, given only a few sampled columns of $\mathbf{M}$. We shall further assume an approximate structure that is captured in the following way: $\mathbf{M} = \mathbf{QS} + \mathbf{E}$. The (known) matrix $\mathbf{S}$ captures the approximate side information, $\mathbf{Q}$ is an unknown coefficient matrix and $\mathbf{E}$ is also unknown and captures the fact that the side information $\mathbf{S}$ is only approximate.

Given the special nature of our problem formulation, it is challenging to find appropriate comparison algorithms. To this end, we consider CUR decompositions for matrix approximation, wherein the original formulations [20], [21] assumed access to both full columns and full rows of the true matrix. We note that CUR+ [19] considers the missing value case and computes an error bound with sample complexity $O(nr \ln r)$, but still employs fully sampled rows and columns as well as additional random samples of the original matrix. A challenge with the CUR family of methods is that they assume access to the true matrix $\mathbf{M}$. We underscore that CUR+ does not match our problem formulation. A perturbed version of CUR (PCUR) [22] assumes access to a noisy version of $\mathbf{M}$ and applies CUR principles for low rank matrix approximation. However, the assumptions of [22] are also mismatched with ours as we do not have access to a full noisy version of the true matrix in our formulation. Additionally, the derivation of sample-complexity bounds for PCUR within our framework is not straightforward. Nonetheless, we will adapt CUR+ and PCUR in order to study the performance of our proposed algorithm. Furthermore, our sampling complexity analysis adopts some techniques of [19].

We also note that while there are other lines of work that are tangentially related to the problem considered herein, none of them can be applied in our setting without non-trivial modifications of the algorithm (and analysis). For example, [23], [24] consider the column sampling mechanism when the true matrix is given, but do not consider matrix completion. Another line of work that studies matrix approximation

strategies is *sketching* [25], [26], where dense and global measurements of the matrix are available, but only projections are employed for the approximation. Finally, [7]–[9] assume uniform random sampling for completion and perfect or imperfect side information.

The key contributions of this work are as follows.

1) We provide a novel problem formulation for low-rank matrix approximation method based on randomly, but fully sampled columns coupled with side information that is captured via a quasi-polynomial structure. An algorithm for this problem is proposed - the Quasi-Polynomial Matrix Approximation (QPMA) algorithm
2) A theoretical spectral bound on the reconstruction error achieved by QPMA is derived. This bound is shown to be only slightly worse, with respect to order of key parameters, to that achievable by prior matrix approximation strategies with significantly weaker assumptions. In particular, Theorem 1, shows that the matrix can be recovered when row space information is provided that is *close* to that of the true matrix.
3) QPMA is compared to CUR+ and PCUR on synthetic data and shown to offer strong performance; furthermore, QPMA is validated on data from the original quantum chemistry application.

While our strategy is motivated by a specific application, we believe our algorithm and analysis have greater applicability to problems wherein side information can be succinctly captured by row space information and there are sampling constraints such as undersampled radial magnetic resonance imaging where low-rank matrix completion has been successfully applied [27], [28]. Furthermore, arbitrary side information can be captured by the matrix $\mathbf{S}$. Characterizing the applicability of our methods to more general problems is an avenue for future work. In our prior work [16], an algorithm for column sampling coupled with quasi-polynomial side information was proposed and numerically shown to offer good performance. A challenge with the proposed algorithm in [16] was controlling the rank of the approximated matrix. With the modified approach herein, we can carefully control for rank while providing theoretical guarantees that are based on algorithm and system parameters.

This paper is organized as follows. Section II introduces the quantum chemistry application that motivates this work and provides the background of the system model. After that, the formal problem setting and the proposed algorithm are provided in Section III. The main result of this paper derives a spectral reconstruction error coupled with the key parameters such as target rank, the true rank and a given polynomial degree. This result is presented in Section IV with the discussions on time complexity and comparison with prior art. The key simulation results to evaluate the main theorem, as well as comparisons with prior art are provided in Section V. The remainder of the paper is the proof of the theorem and key lemmas.

The following notation is adopted herein. We define $[m] \doteq \{1, \ldots, m\}$. Bold upper case letters $\mathbf{M}$ denote matrices. For a given set $\mathcal{C}$, $|\mathcal{C}|$ denotes the cardinality of the set. We use $\|\cdot\|$ to denote the spectral norm unless otherwise specified. We use $\overset{\text{SVD}}{=}$, $\overset{r-\text{SVD}}{=}$ to denote the singular value decomposition (SVD) and the reduced (rank-r) SVD of a matrix, respectively. $\sigma_r(\mathbf{M})$ refers to the r-th largest singular value of a matrix $\mathbf{M}$. $\mathbf{M}^\dagger$ denotes the pseudo-inverse of $\mathbf{M}$. Finally, throughout this paper, with a slight abuse of terminology, we use the terms column and row space of a matrix, $\mathbf{M}$ to mean the *best* r-*dimensional* approximation for the respective spaces. We use the order notation $O(\cdot)$ to show the asymptotic dependence with respect to data dimensions.

II. Motivating Application

We provide the motivation for this work and its applications. In the study of chemical reactions using quantum chemistry methods, Variational Transition State Theory (VTST) is a technique for calculating reaction rate coefficients that describe kinetics [29]. VTST suffers from high computational cost as it requires the calculation of expensive quantum mechanical Hessians of energy at several points constituting the minimum energy path (MEP) of a reaction. Prior efforts towards reducing computational effort include interpolated VTST (I-VTST), which fits splines under tension to energies, gradients, and Hessians calculated at arbitrary points on the MEP [30]. In our prior work [14], [15], we showed that randomized sampling coupled with an algebraic variety constraint [31] could accurately complete an incomplete matrix of Hessian eigenvalues constituting the MEP when only a small, randomly sampled set of elements are available. In particular, the algebraic variety constraint is well-matched to this problem as, within the reaction path Hamiltonian (RPH) framework [32], the harmonic potential energy terms $V(s, \mathbf{q})$ are formulated into a polynomial expression of the eigenvalues $\{w_k^2\}$ of Hessian matrix and displacements along vibrational normal modes $\{q_k^2\}$ as

$$V(s, \mathbf{q}) = V_0(s) + \sum_{k=1}^{n} w_k^2 q_k^2, \tag{1}$$

where $n = 3n_a - 7$ indicates the number of vibrational modes that are orthogonal to the reaction coordinate, n_a is the number of atoms and $V_0(s)$ is the potential energy at a point s on the MEP.

While our algorithm proposed in [14] was computationally efficient and provided a proof-of-concept, it assumed randomized sampling, whereas, pragmatically one can compute one Hessian at a time, which corresponds to one column of the true matrix.

The true matrix $\mathbf{M}$ of Hessian eigenvalues constituting the MEP is constructed by the potential energy term of the reaction path Hamiltonian [32], [33]. The true matrix $\mathbf{M}$ is given by

$$\mathbf{M} \in \mathbb{R}^{n \times m} = \begin{bmatrix} \omega_1^2(s_1) & \omega_1^2(s_2) & \ldots & \omega_1^2(s_m) \\ \omega_2^2(s_1) & \omega_2^2(s_2) & \ldots & \omega_2^2(s_m) \\ \vdots & & & \\ \omega_n^2(s_1) & \omega_n^2(s_2) & \ldots & \omega_n^2(s_m) \end{bmatrix}.$$

$\{\omega_i\}, i \in [n]$, constitutes the set of vibrational frequencies of the system obtained upon projecting out the reaction coordinate, translations, and rotations from the Hessian. Each column $\mathbf{M}_j, j \in [m]$, is comprised of n eigenvalues $\{w_i^2\}, i \in [n]$,

of the projected quantum mechanical Hessian matrix. The reaction coordinate is parameterized by s_i, where $i \in [m]$, defined to be zero at the transition state, negative in the reactant region (with reactant represented by s_1), and positive in the product region (with product represented by s_m). The goal is to approximate $\mathbf{M}$ given a few full columns in a way that VTST rate coefficients can be estimated with reasonable accuracy.

III. Problem Formulation and Algorithm

We next present the concrete problem formulation, the proposed optimization strategy, and finally our main guarantee.

A. Problem setting

Let $\mathbf{M} \in \mathbb{R}^{n\times m}$ be the true matrix of rank k. In this paper, we consider the problem of obtaining a rank-r approximation of $\mathbf{M}$ from d randomly sampled columns. In particular, we consider the following regime

$$\underbrace{r}_{\text{target rank}} \leq \underbrace{d}_{\text{\# of sampled columns}} \leq \underbrace{k}_{\text{true rank}}$$

In contrast to traditional Matrix Completion, we seek a lower rank approximation of the matrix (w.r.t. the true rank). This allows us to obtain our main guarantees even when the number of columns sampled is smaller than the actual rank.

To describe the chemical rate reaction processes, as mentioned in Section II, given a list of reaction coordinate values, $\mathbf{s} = [s_1, \ldots, s_m]$ and polynomial order l we model $\mathbf{M}$ as

$$\mathbf{M} = \mathbf{QS} + \mathbf{E} \tag{2}$$

where, $\mathbf{Q} \in \mathbb{R}^{n\times l}$ is an unknown polynomial coefficient matrix, the structural side information matrix, $\mathbf{S} \in \mathbb{R}^{l\times m}$ encodes the known polynomial information (described next) and $\mathbf{E}$ is the perturbation/noise matrix.

Per Section II, the eigenvalues of the Hessian matrix $\mathbf{M}$ is quasi-polynomial, we assume that the side information $\mathbf{S}$ has the following structure

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ s_1 & s_2 & \ldots & s_m \\ s_1^2 & s_2^2 & \ldots & s_m^2 \\ \vdots & \vdots & \vdots & \vdots \\ s_1^{l-1} & s_2^{l-1} & \ldots & s_m^{l-1} \end{bmatrix}. \tag{3}$$

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, $\mathbf{A}$, can be equivalently expressed as

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

Table I summarizes the parameters for the introduced matrices.

TABLE I
Description of various matrix ranks

Matrix	Rank	Relationship
$\mathbf{M}$	k	$r \leq l, d \leq k$
$\mathbf{QS}$	l	
$\mathbf{A} = \mathbf{M\Psi}$	$\leq d$	

Before setting up the optimization problem, we define the following quantities. We denote the SVD of the true matrix, $\mathbf{M}$ as

$$\mathbf{M} \overset{\text{SVD}}{=} \mathbf{U\Sigma V}^{\mathsf{T}} = \underbrace{\mathbf{U}_M\mathbf{\Sigma}_M\mathbf{V}_M^{\mathsf{T}}}_{\text{rank-}r\text{-approximation}} + \underbrace{\mathbf{U}_{M,\perp}\mathbf{\Sigma}_{M,\perp}\mathbf{V}_{M,\perp}^{\mathsf{T}}}_{\text{remainder}} \tag{4}$$

Notice that when the target rank r is smaller than the true rank k, the second term above is non-zero. Similarly, we define the SVD of $\mathbf{QS}$ as

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

B. Quasi-Polynomial Matrix Approximation Algorithm

We next introduce the proposed optimization strategy, Quasi Polynomial Matrix Approximation (QPMA). Note that if the column and row space information of $\mathbf{M}$, i.e., $\mathbf{U}_M$ and $\mathbf{V}_M$ respectively, were known, a natural way to cast the optimization that takes into account the structural information including the desired rank r approximation of $\mathbf{M}$ is as

$$\min_{\mathbf{Z}} \ \left\|\mathbf{A} - \mathbf{U}_M\mathbf{Z}\mathbf{V}_M^{\mathsf{T}}\mathbf{\Psi}\right\|_F^2 \tag{6}$$

However, since we do not know $\mathbf{U}_M$ and $\mathbf{V}_M$, we need to estimate them using the prior structural information of $\mathbf{M}$. To this end, the proposed QPMA algorithm is comprised of three stages: (i) estimating the column space of $\mathbf{M}$; (ii) followed by estimating the unknown polynomial coefficient matrix, $\mathbf{Q}$, and subsequently estimating the row space of $\mathbf{M}$ by leveraging the quasi polynomial structure; and (iii) the final matrix approximation step constrained to the row and column space approximations obtained previously. The complete algorithm is summarized in Algorithm 1 (QPMA).

We first estimate the column space of $\mathbf{M}$ using $\mathbf{A} = \mathbf{M\Psi}$. We argue that as long as enough independent columns are sampled (this is shown in Lemma 1), the following optimization gives us a good estimate

$$\mathbf{U}_A = \underset{\tilde{\mathbf{U}}\in\mathbb{R}^{n\times r}, \tilde{\mathbf{U}}^{\mathsf{T}}\tilde{\mathbf{U}}=\mathbf{I}}{\operatorname{argmin}} \left\|(\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^{\mathsf{T}})\mathbf{A}\mathbf{A}^{\mathsf{T}}\right\|_2.$$

[1] For example, when $m = 4$, $\mathcal{C} = \{1, 3, 4\}$, i.e., $d = 3$,

$$\mathbf{\Psi} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

From Eckart-Young-Mirsky theorem, the solution to the above is given by the rank-r SVD of $\mathbf{A}$ (the matrix formed by the left singular vectors corresponding to the top-r singular values),

$$\mathbf{A} \overset{\text{r-SVD}}{=} \mathbf{U}_A \mathbf{\Sigma}_A \mathbf{V}_A^{\mathsf{T}}. \tag{7}$$

Algorithm 1 Quasi-Polynomial Matrix Approximation

Input: $\mathbf{A} \in \mathbb{R}^{n \times d}$(A matrix of sampled columns), $\mathbf{S} \in \mathbb{R}^{l \times m}$(A polynomial basis matrix), $\mathbf{\Psi}$ (Column sampling operator)
Parameters: A target rank r, Degree of polynomial l, Step size η, Max iteration T
Initialization: generate each entry of $\hat{\mathbf{Q}}_1$ independently from $\mathcal{N}(0,1)$
Column space estimation
Do rank-r SVD of $\mathbf{A}$ as $\mathbf{A} \overset{r-\text{SVD}}{=} \mathbf{U}_A \mathbf{\Lambda}_A \mathbf{V}_A^{\mathsf{T}}$
Row space estimation
For $t \in [T]$, do
$\hat{\mathbf{Q}}_{t+1} = \hat{\mathbf{Q}}_t - \eta \left(\mathbf{A} - \hat{\mathbf{Q}}_t \mathbf{S}\mathbf{\Psi} \right) (\mathbf{S}\mathbf{\Psi})^{\mathsf{T}}$
With $\hat{\mathbf{Q}} \equiv \hat{\mathbf{Q}}_{T+1}$, compute $\hat{\mathbf{Q}}\mathbf{S} \overset{r-\text{SVD}}{=} \hat{\mathbf{U}}_{QS} \hat{\mathbf{\Lambda}}_{QS} \hat{\mathbf{V}}_{QS}^{\mathsf{T}}$
Matrix approximation
Using $\mathbf{U}_A$ and $\hat{\mathbf{V}}_{QS}$, for $t \in [T]$, do
$\hat{\mathbf{Z}}_{t+1} = \hat{\mathbf{Z}}_t - \eta \mathbf{U}_A^{\mathsf{T}} \left(\mathbf{A} - \mathbf{U}_A \hat{\mathbf{Z}}_t \hat{\mathbf{V}}_{QS}^{\mathsf{T}} \mathbf{\Psi} \right) \left(\hat{\mathbf{V}}_{QS}^{\mathsf{T}} \mathbf{\Psi} \right)^{\mathsf{T}}$
Obtain $\hat{\mathbf{Z}} = \hat{\mathbf{Z}}_{T+1}$
Complete $\hat{\mathbf{M}} = \mathbf{U}_A \hat{\mathbf{Z}} \hat{\mathbf{V}}_{QS}^{\mathsf{T}}$
Output: $\hat{\mathbf{M}} = \mathbf{U}_A \hat{\mathbf{Z}} \hat{\mathbf{V}}_{QS}^{\mathsf{T}}$.

We next estimate the unknown polynomial coefficient matrix, $\mathbf{Q}$ as follows

$$\hat{\mathbf{Q}} = \underset{\tilde{\mathbf{Q}}}{\operatorname{argmin}} \ \left\| \mathbf{A} - \tilde{\mathbf{Q}} \mathbf{S}\mathbf{\Psi} \right\|_F^2. \tag{8}$$

This is a standard regression problem that admits a closed form solution, but it is computationally expensive to compute a pseudo-inverses. Thus, we instead consider a gradient descent approach [34]. Concretely, we define $g(\mathbf{Q}) = \|\mathbf{A} - \mathbf{Q}\mathbf{S}\mathbf{\Psi}\|_F^2$. The gradient of $g(\mathbf{Q})$ with respect to $\mathbf{Q}$ is given by

$$\nabla_{\mathbf{Q}} g(\mathbf{Q}) = 2 \left(\mathbf{A} - \mathbf{Q}\mathbf{S}\mathbf{\Psi} \right) (\mathbf{S}\mathbf{\Psi})^{\mathsf{T}}.$$

We then repeat the following update rule at each iteration $t = [T]$ until convergence:

$$\hat{\mathbf{Q}}_{t+1} = \hat{\mathbf{Q}}_t - \eta \left(\mathbf{A} - \hat{\mathbf{Q}}_t \mathbf{S}\mathbf{\Psi} \right) (\mathbf{S}\mathbf{\Psi})^{\mathsf{T}}, \tag{9}$$

where η is an appropriately chosen step size. Now if $\hat{\mathbf{Q}} \equiv \hat{\mathbf{Q}}_T$ is a good approximation of $\mathbf{Q}$ (this is shown in Lemma 2), we can obtain the row space information of $\mathbf{M}$ through the following minimization

$$\hat{\mathbf{V}}_{QS} = \underset{\tilde{\mathbf{V}} \in \mathbb{R}^{m \times r}, \tilde{\mathbf{V}}^{\mathsf{T}} \tilde{\mathbf{V}} = \mathbf{I}}{\operatorname{argmin}} \left\| (\mathbf{I} - \tilde{\mathbf{V}} \tilde{\mathbf{V}}^{\mathsf{T}}) (\hat{\mathbf{Q}}\mathbf{S})^{\mathsf{T}} (\hat{\mathbf{Q}}\mathbf{S}) \right\|_2 \tag{10}$$

Again, the solution to the above is readily obtained through a rank-r SVD of $\hat{\mathbf{Q}}\mathbf{S}$.

Finally, we exploit the row and column space estimates to obtain the low-rank approximation as follows

$$\hat{\mathbf{Z}} = \arg\min_{\tilde{\mathbf{Z}}} \ \left\| \mathbf{A} - \mathbf{U}_A \tilde{\mathbf{Z}} \hat{\mathbf{V}}_{QS}^{\mathsf{T}} \mathbf{\Psi} \right\|_F^2. \tag{11}$$

We observe that (11) is also a regression problem that we solve through an gradient descent method. Define

$$f(\mathbf{Z}) = \left\| \mathbf{A} - \mathbf{U}_A \mathbf{Z} \hat{\mathbf{V}}_{QS}^{\mathsf{T}} \mathbf{\Psi} \right\|_F^2. \tag{12}$$

The gradient of $f(\mathbf{Z})$ is given by

$$\nabla_{\mathbf{Z}} f(\mathbf{Z}) = 2 \mathbf{U}_A^{\mathsf{T}} \left(\mathbf{A} - \mathbf{U}_A \mathbf{Z} \hat{\mathbf{V}}_{QS}^{\mathsf{T}} \mathbf{\Psi} \right) \left(\hat{\mathbf{V}}_{QS}^{\mathsf{T}} \mathbf{\Psi} \right)^{\mathsf{T}},$$

This finally yields the reconstructed matrix

$$\hat{\mathbf{M}} = \mathbf{U}_A \hat{\mathbf{Z}} \hat{\mathbf{V}}_{QS}^{\mathsf{T}}, \tag{13}$$

This concludes the algorithm.

IV. Main Result and Proof Sketch

In this section, we provide our main result and the proof sketch. We require the following definitions before presenting the main result. We consider the following standard definition of matrix incoherence [2].

Definition 1 (Incoherence)**.** *Let $\mathbf{X}$ be a $n \times m$ matrix of rank r and $\mathbf{X} \overset{r\text{-}SVD}{=} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$. 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*

$$\mu(\mathbf{X}) = \max \left(\max_{i \in [n]} \frac{n}{r} \|\mathbf{u}_i\|_2^2, \max_{j \in [m]} \frac{m}{r} \|\mathbf{v}_j\|_2^2 \right).$$

Incoherence is a necessary assumption to ensure the "energy" is spread out uniformly to complete a matrix from a few randomly chosen entries. We note that despite the fact that our work deals with the setting wherein a few randomly chosen *columns* are observed (as opposed to a few randomly chosen *entries* that standard MC studies), the inclusion of the quasi-polynomial side information allows us to work with the standard incoherence definition. In our analysis, we use the shorthand notation, $\mu \doteq \mu(\mathbf{M})$ and $\hat{\mu} \doteq \mu(\hat{\mathbf{M}})$ [2].

Next, we review strong convexity of a function [34, sec 3].

Definition 2 (Strong Convexity)**.** *A differentiable[3] function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with parameter $\alpha > 0$ if the following holds for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$,*

$$f(\mathbf{x}) \geq f(\mathbf{x}') + \nabla f(\mathbf{x}')(\mathbf{x} - \mathbf{x}') + \frac{\alpha}{2} \|\mathbf{x} - \mathbf{x}'\|_2^2.$$

We use Definition 2 to derive convergence guarantees for the row space estimation, and the matrix approximation steps of QPMA.

[2] In our current result, we assume that the output of QPMA (Algorithm 1) is incoherent. We will consider eliminating this assumption as part of future work.

[3] If f is non-differentiable, then the gradient of f is replaced by its sub-gradient.

A. Main Result

We need the following assumption before presenting our main result.

Assumption 1. *Let* $\delta_1 := \sigma_r(\mathbf{M}-\mathbf{E}) - \sigma_{r+1}(\mathbf{M}) > 0$.

Assumption 2. *Let* $\delta_2 := |\sigma_r(\mathbf{QS}) - \sigma_{r+1}(\hat{\mathbf{Q}}\mathbf{S})| > 0$.

Observe that δ_1 captures the effective eigengap of $\mathbf{M}$. In contrast, δ_2 captures the effective eigengap of estimated side information $\mathbf{QS}$. We provide a numerical validation of the fact

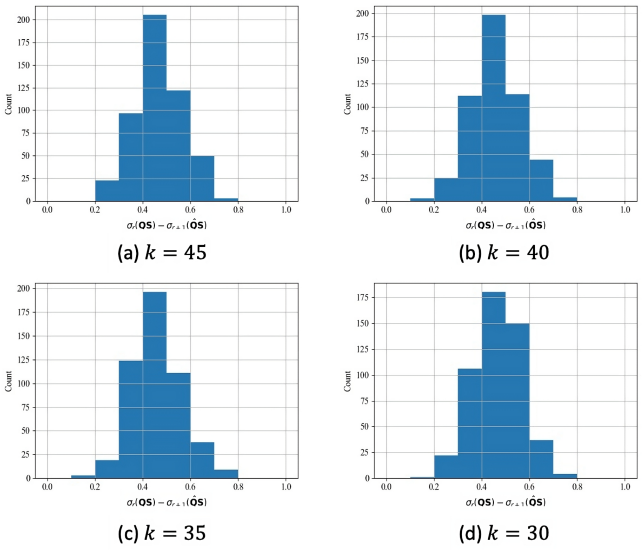

Fig. 2. Comparison of $|\sigma_r(\mathbf{QS}) - \sigma_{r+1}(\hat{\mathbf{Q}}\mathbf{S})|$ for different values of k. The histogram is plotted over 300 independent realizations of the data. Further details are provided in the appendix.

that typically $\delta_2 > 0$ in Fig.2. We also provide more numerical experiments for the same in the Appendix. We now present our main result.

Theorem 1. *Consider measurements that satisfy Assumption 1. Assume that d columns are sampled uniformly at random from the underlying ground truth, $\mathbf{M}$. Then, if $d \geq \max\left\{c_1\mu r \ln r, c_2\hat{\mu}^2 r^2 \ln r\right\}$, with probability at least $1 - c_3 r^{-10}$ we have*

$$\frac{\|\mathbf{M}-\hat{\mathbf{M}}\|_2^2}{\|\mathbf{M}\|_2^2} \leq 4\frac{\sigma_{r+1}^2(\mathbf{M})}{\sigma_1^2(\mathbf{M})}\left(1+\frac{4m}{d}\right)\left(3+\frac{n}{d}\right) + 64\|\mathbf{E}\|_F^2\left(1+\frac{4m}{d}\right)\left(\frac{1}{\delta_1^2}+\frac{1}{\delta_s^2}\right) \tag{14}$$

where $\delta_s := \frac{\delta_2}{\|(\mathbf{S\Psi})^\dagger\mathbf{S}\|_F}$ and $c_1, c_2, c_3 > 0$ are numerical constants.

Proof. Theorem 1 is proved in the Appendix. The proof follows from applying large-deviation style results from random matrix theory [35] to ensure that the loss-function in (12) is well-behaved as long as we sample a sufficient number of columns, followed by a careful application of Wedin's theorem [36]. ⊠

If $\mathbf{E} = \mathbf{0}$, we have the following Corollary.

Corollary 1 (Perfect Side-Information)**.** *Under the conditions of Theorem 1, if $\mathbf{E} = \mathbf{0}$, then with probability at least $1 - Cr^{-10}$, where $C > 0$,*

$$\left\|\mathbf{M}-\hat{\mathbf{M}}\right\|_2^2 \lesssim \frac{80mn}{d^2}\sigma_{r+1}^2(\mathbf{M}) = O\left(\frac{mn}{d^2}\|\mathbf{M}-\mathbf{M}_r\|_2^2\right) \tag{15}$$

where $\mathbf{M}_r$ is the best rank-r approximation of $\mathbf{M}$.

B. Discussion

Interpreting the Signal-to-Noise Ratio. Recall from Assumption 1 that δ_1 captures the effective eigengap of $\mathbf{M}$ and thus a natural interpretation of the term $\frac{\delta_1^2}{\|\mathbf{E}\|_F^2}$ is the "signal-to-noise ratio"(SNR). Furthermore, we observe from Theorem 1 that $\delta_s := \frac{\delta_2}{\|(\mathbf{S\Psi})^\dagger\mathbf{S}\|_F}$ essentially measures *how informative* the side information, $\mathbf{S}$ is. More precisely, first consider δ_s: Notice that the larger the numerator term, i.e., the effective singular value gap, $|\sigma_r(\mathbf{QS}) - \sigma_{r+1}(\hat{\mathbf{Q}}\mathbf{S})|$, the more informative, the side-information is. The denominator term, $\|(\mathbf{S\Psi})^\dagger\mathbf{S}\|$, on the other hand, measures how much of the side information is effectively captured after the column-sampling process. Observe that if $\mathbf{\Psi} = \mathbf{I}$, then $\|(\mathbf{S\Psi})^\dagger\mathbf{S}\|_F = \|(\mathbf{S})^\dagger\mathbf{S}\|_F = \sqrt{l}$, and as expected, this value reduces as the number of sampled columns, d, reduces. Finally, we emphasize that from the perspective of the motivating application, we can control $\mathbf{S}$ and thus, it is possible to ensure that $\|(\mathbf{S\Psi})^\dagger\mathbf{S}\|_F = O(1)$. Finally, consider δ_2: without further assumptions on the data it is not possible to obtain a bound, in general; however, it is reasonable to assume that $\delta_2 = O(1)$ (with respect to n, r). We provide a numerical validation for this point in Fig. 2 and a more exhaustive evaluation in the Appendix. Additionally, through our simulations we ascertained that this (theoretical) dependence on δ_2 is weak, i.e., the final error does not decay quadratically with δ_2.

Finally, with a slight misuse of terminology, we use

$$\text{SNR} := \frac{1}{\|\mathbf{E}\|_F}\left(\frac{1}{\delta_1^2}+\frac{1}{\delta_s^2}\right)^{-1/2}$$

as a measure of the effective signal-to-noise ratio in the sequel.

In Theorem 1, we focus on the two sources of error: (i) the *unrecoverable energy* that arises due to fact that the original matrix is high-rank; and (ii) the *imperfect side-information*. The first term in (14) represents the unrecoverable energy, as we seek a low-rank approximation of a high-rank matrix. Even if we had perfect side information, i.e., $\mathbf{E} = \mathbf{0}$ there will be an error incurred due to the low-rank approximation. We also observe that QPMA suffers a multiplicative factor of $O(m/d)$ coupled with the best rank-r approximation error, $\|\mathbf{M}-\mathbf{M}_r\|_2 = \sigma_{r+1}(\mathbf{M})$. This is owing to the fact that we solve a harder problem than classical rank-r approximation and this multiplicative factor is standard in the high-rank matrix approximation literature [21], [25], [37], [38]. The second term in (14) occurs due to the imperfect nature of the side information, i.e., since $\mathbf{E} \neq \mathbf{0}$. We emphasize that since our main result does not assume any statistical or generative models on the noise, it is highly non-trivial to make further

deductions. Thus, we consider specific noise models, and the side-information matrices in future work.

Comparison with CUR+ [19]. We assume for the rest of the paper that the incoherences, are constant[4], i.e., $\mu, \hat{\mu} = O(1)$. With this, it is easy to see from Theorem 1 that $d = O(r^2 \ln r)$. Observe that in order to obtain a non-trivial rank r approximation, one needs to sample at least r columns of $\mathbf{M}$ even with perfect side information. Theorem 1 shows that with mismatched side-information and unstructured noise, QPMA obtains a good approximation with just $O(r^2 \ln r)$ columns. We contrast with CUR+ since its sampling structure is the most similar to our problem setting and is also the state-of-the-art in high-rank matrix approximation with incomplete measurements. As opposed to QPMA, CUR+ requires a $d = O(r \ln r)$ randomly chosen rows and columns, and an additional $O(r^2 \ln r)$ randomly chosen entries. Thus, by imposing a significantly weaker assumption: a quasi-polynomial side information instead of observing a subset of rows, QPMA attains a sample complexity bound that is a factor of r worse than that of CUR+. We believe that this bound can be improved by a more refined proof technique for Lemma 2 which we defer to future work. Finally, we note that the weaker set of assumptions considered in this work come at a cost: unlike the results of CUR+ which require stronger assumptions, our Theorem 1 shows a dependence on $\hat{\mu}, \delta_2$ and δ_s (these are assumed to be $O(1)$). However, extensive numerical experiments suggest that these dependencies are loose. Further examination of these issues will be undertaken in future work.

Interpreting the Error. Many prior error analyses for high-rank matrix approximation [20], [21] have the following common structure for the error bound:

$$\left\|\mathbf{M} - \hat{\mathbf{M}}\right\| \leq (1+\epsilon_1)\left\|\mathbf{M} - \mathbf{M}_r\right\| + \epsilon_2 \left\|\mathbf{M}\right\|, \tag{16}$$

where $\mathbf{M}_r$ is the best rank-r approximation of a matrix $\mathbf{M}$ and $\epsilon_1 > 0$, and $\epsilon_2 \in (0,1)$ are derived constants that are specialized to the problem. We see that we can formulate our error bound from Theorem 1 in a similar fashion,

$$\left\|\mathbf{M} - \hat{\mathbf{M}}\right\| \leq O\left(\frac{m}{d}\left\|\mathbf{M} - \mathbf{M}_r\right\|\right) + O\left(\sqrt{\frac{m}{d}}\mathrm{SNR}^{-1}\|\mathbf{M}\|\right)$$

where, without loss of generality, we assume $m \geq n$. As previously mentioned, the scaling factor $\frac{m}{d} (\equiv 1+\epsilon_1)$ is, in general, unavoidable due the high-rank nature of the true matrix in addition to sub-sampling of the columns. We also notice that $\epsilon_2 = C\sqrt{\frac{m}{d}} \cdot \frac{1}{\mathrm{SNR}}$, i.e., the noise is amplified by the square root of the sub-sampling factor, $\sqrt{\frac{m}{d}}$ as well. We emphasize that unlike results in PCA, wherein there is a "denoising" effect with increasing the number of observations, matrix approximation algorithms do not possess the ability to denoise, the observations. However, as expected, increasing the number of observed columns reduces the approximation error and approaches the best-case scenario of $C \cdot \mathrm{SNR}^{-1}$ as $m \to d$. Finally, we mention that in the setting where $\|\mathbf{E}\|_F = 0$, and if the matrix is "roughly square", i.e., $m = O(n)$, our result improves upon CUR+ [19, Theorem 2] by a factor of $\sqrt{m}$.

[4]In this paper, we use the order notation with respect to m, n.

Time complexity of QPMA. We next derive the computational complexity of QPMA (Algorithm 1). The column space $\mathbf{U}_A$ is estimated through a rank-r SVD on $\mathbf{A}$, and this takes $O(ndr)$ time [39][5]. Next, the row space estimation step is performed by first estimating the polynomial coefficient matrix $\mathbf{Q}$ by gradient descent (GD). The run time for the corresponding matrix multiply in each iteration is $O(\max(nl^2d, nd^2l)) = O(nd^2l)$ (we assume $d > l$ without loss of generality) and thus the overall complexity of GD is $O(nd^2l)$ given a bounded gradient assumption and a bounded initial error[6] [34, Sec 9]. Next, the rank-r SVD of $\hat{\mathbf{Q}}\mathbf{S}$ can be performed in $O(nmr)$ time. Finally, the per-iteration complexity for the matrix approximation step is $O(\max(n^2dr, nd^2r)) = O(n^2dr)$ and since we assume that the number of iterations, $T = O(1)$, the overall running time for GD is $O(n^2dr)$. Thus, the overall computational complexity of QPMA is $O(\max(nmr, n^2dl))$ that is equal (up to constant factors) to performing a rank-r SVD on the original matrix, $\mathbf{M}$.

C. Proof Sketch and key Lemmas

Here we provide the proof sketch and the main Lemmas required to prove Theorem 1. The complete proofs are provided in the Appendix.

We first bound the error as $\|\mathbf{M} - \hat{\mathbf{M}}\|_2^2$ as

$$\begin{aligned}
\left\|\mathbf{M} - \hat{\mathbf{M}}\right\|_2^2 &= \left\|\mathbf{M} - \mathbf{U}_A\hat{\mathbf{Z}}\hat{\mathbf{V}}_{QS}^{\mathsf{T}}\right\|_2^2 \\
&= \|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}} \\
&\qquad + \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}} - \mathbf{U}_A\hat{\mathbf{Z}}\hat{\mathbf{V}}_{QS}^{\mathsf{T}}\|_2^2 \\
&\overset{(a)}{\leq} 2\underbrace{\left\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2}_{\odot} \\
&\qquad + 2\underbrace{\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}} - \mathbf{U}_A\hat{\mathbf{Z}}\hat{\mathbf{V}}_{QS}^{\mathsf{T}}\right\|_2^2}_{\circledcirc}
\end{aligned} \tag{17}$$

where (a) follows from the triangle inequality and the fact that for $a, b \geq 0$, $(a+b)^2 \leq 2(a^2+b^2)$. Next, recall that $\mathbf{P}_{\mathbf{U}_A} = \mathbf{U}_A\mathbf{U}_A^{\mathsf{T}}$ and $\mathbf{P}_{\hat{\mathbf{V}}_{QS}} = \hat{\mathbf{V}}_{QS}\hat{\mathbf{V}}_{QS}^{\mathsf{T}}$. Notice that can obtain high probability bounds on $\odot$ and $\circledcirc$, we are done. To that end, we first consider $\odot$.

Note that $\odot$ captures the energy of the true matrix, $\mathbf{M}$ orthogonal to the estimated (r-dimensional) row and column spaces. We provide a bound for this below in Lemma 1.

Lemma 1. *Consider measurements that satisfy Assumption 1. Then, if $d \geq c_1\mu r \ln r$, under the conditions of Theorem 1,*

[5]Note that we only require the top-r singular vectors, but do not require the singular values, and hence there is no dependence on the singular value gap

[6]In this paper, we assume that the number of iterations for the GD step is $O(1)$. We do this since without additional statistical assumptions on the signal model, characterizing T is very complex, and beyond the scope of this paper

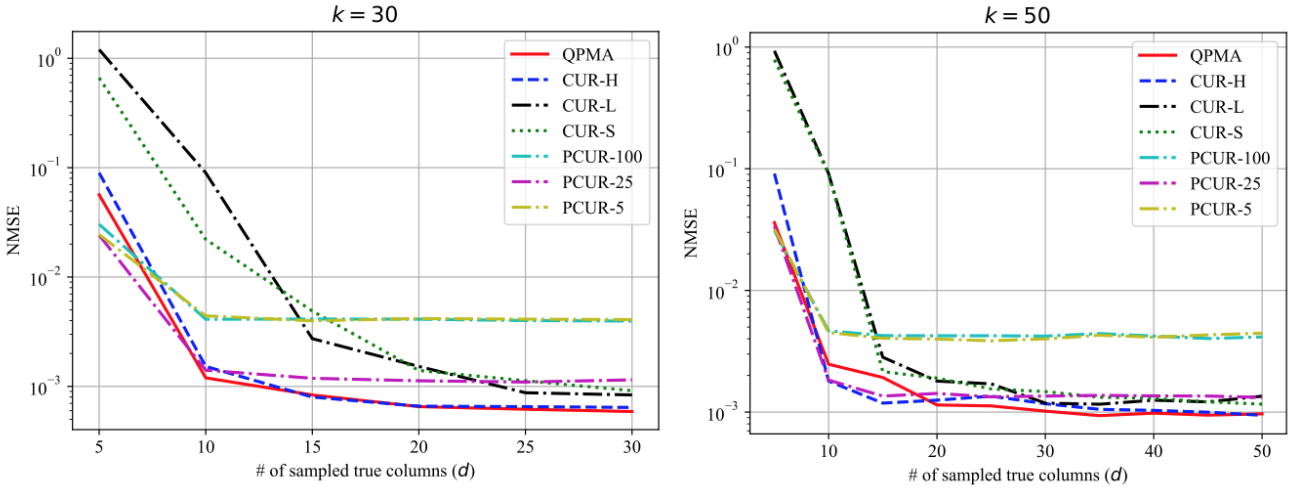

Fig. 3. The NMSE versus the number of true sampled columns, d. Here, $k = \{30, 50\}$ and $l = 5$.

with probability at least $1 - c_2 r^{-10}$ *we have that*

$$\left\| \mathbf{M} - \mathbf{P}_{\mathbf{U}_A} \mathbf{M} \mathbf{P}_{\hat{\mathbf{V}}_{QS}} \right\|_2^2 \leq 2\sigma_{r+1}^2(\mathbf{M}) \left(3 + \frac{n}{d}\right) + 32\sigma_1^2(\mathbf{M}) \|\mathbf{E}\|_F^2 \left(\frac{1}{\delta_1^2} + \frac{1}{\delta_s^2}\right) \tag{18}$$

where $\delta_s := \frac{\delta_2}{\|(\mathbf{S}\Psi)^\dagger \mathbf{S}\|_F}$ *and* $c_1, c_2 > 0$ *are numerical constants.*

The complete proof of Lemma 1 is provided in Appendix B. The proof follows from first invoking [40, Theorem 6] to bound the energy of $\mathbf{M}$ orthogonal to $\mathbf{U}_A$ and $\hat{\mathbf{V}}_{QS}$, where $\hat{\mathbf{V}}_{QS}$ is the row space of $\hat{\mathbf{Q}}\mathbf{S}$, followed by a careful application of Wedin's Theorem [36] (provided in appendix as Theorem 3) to bound the "distance" between the "true row space" $\mathbf{V}_{QS}$ defined in (5) and the estimated $\hat{\mathbf{V}}_{QS}$ obtained from (10). These are provided as Lemmas 4 and 5 respectively.

Akin to the result of [19, Theorem 2] and as explained previously, Lemma 1, consists of error due to the fact that $\text{rank}(\mathbf{M}) = k \gg r$ and sub-sampling of columns (both contribute to the first term). When $k = r$, the first term is zero since $\sigma_{r+1}(\mathbf{M}) = 0$. The second term corresponds to the error due to noise and imperfect nature of side information.

Next, $\circledcirc$ essentially captures the error in the final matrix approximation step, estimation of $\mathbf{Z}$. This is bounded using Lemma 2 below.

Lemma 2. *Consider measurements that satisfy Assumption 1. Let the objective function in* (6) *be* α*-strongly convex with* $\alpha \geq d/2m$*. Then, if* $d \geq c_2 \hat{\mu}^2 r^2 \ln r$*, under the conditions of Theorem 1, with probability at least* $1 - r^{-c_1}$ *we have that*

$$\|\mathbf{P}_{\mathbf{U}_A} \mathbf{M} \mathbf{P}_{\hat{\mathbf{V}}_{QS}} - \mathbf{U}_A \hat{\mathbf{Z}} \hat{\mathbf{V}}_{QS}^{\mathsf{T}}\|_2^2 \leq \frac{4m}{d} \left[2\sigma_{r+1}^2(\mathbf{M}) \left(3 + \frac{n}{d}\right) + 32\sigma_1^2(\mathbf{M}) \|\mathbf{E}\|_F^2 \left(\frac{1}{\delta_1^2} + \frac{1}{\delta_s^2}\right) \right],$$

where $\delta_s := \frac{\delta_2}{\|(\mathbf{S}\Psi)^\dagger \mathbf{S}\|_F}$ *and* $c_1, c_2 > 0$ *are numerical constants.*

The proof of Lemma 2 is provided in Appendix A-B and the proof follows by leveraging the fact that the objective function, $f(\mathbf{Z})$, in (12) is α-strongly convex with $\alpha \geq d/2m$, the result of Lemma 3 and some simple linear algebra tricks.

We next show that the objective function in (12), $f(\mathbf{Z})$ is indeed strongly convex with the requisite parameter setting in Lemma 3.

Lemma 3. *Under the conditions of Theorem 1, with probability at least* $1 - r^{-c_1}$*, the objective function,* $f(\mathbf{Z})$ *in* (12)*, is* α*-strongly convex with* $\alpha \geq \frac{d}{2m}$ *as long as*

$$d \geq c_2 \hat{\mu}^2 r^2 \ln r.$$

where $c_1, c_2 > 0$ *are numerical constants.*

Intuitively, Lemma 3 shows that as long as the number of columns is large enough, the objective function for gradient descent has a quadratic lower bound on the curvature. The proof follows a careful application of a large-deviation result for sums of random matrices [35] followed by linear algebraic computation.

Combining Lemmas 1, 2 and 3 completes the proof.

V. Numerical Results

Herein, we investigate QPMA's performance on both synthetic and real-world data. All experiments on synthetic data are averaged over 100 independent iterations. The code can be found at https://github.com/JeongminChae/QPMA.

Benchmark Algorithms. As noted in Section I, a challenge in finding comparison strategies is that, matrix approximation algorithms typically require an access to the full, **true** row and columns.To this end, we consider CUR+ and PCUR as comparison algorithms for QPMA based on two criteria; (i) different sampling strategies and (ii) informativeness of side information.

CUR+ samples a subset of the full columns and full rows, as well as additional random entries. Thus, for CUR+, the matrix approximation step (corresponding to line 9 in Algorithm 1) is performed with the column space and row space estimated

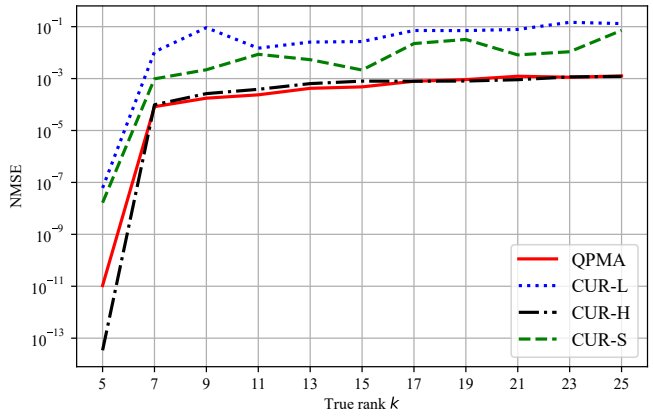

Fig. 4. The NMSE versus the true rank k of $\mathbf{M}$ with fixed $d = 10$, $l = 5$ and $r = 5$.

with the **true** columns and rows, whereas QPMA **estimates** the row space through the side information. We consider three variants of CUR+ [19] which are based on different sampling strategies as outlined in Table II. QPMA samples only d columns, and thus has access to nd entries in total, and CUR-S has access to the same number of entries as QPMA with column, row and random sampling. CUR-L uses fewer total entries with $\frac{d}{2}$ rows and columns each, but no randomized sampling for additional entries. Finally, CUR-H uses an increased number of samples relative to QPMA with d rows and columns each and also does not consider additional randomized sampling of the entries.

TABLE II
COMPARISON OF SAMPLING SCHEMES FOR BASELINE ALGORITHMS

Algorithm	# rows	# columns	random entries	Total samples
CUR-L	$d/2$	$d/2$	0	$nd - d^2/4$
CUR-S	$d/2$	$d/2$	$d^2/4$	nd
CUR-H	d	d	0	$2nd - d^2$
QPMA	0	d	0	nd

Conventional PCUR [22] computes the low-rank matrix of $\mathbf{M}$ from its perturbed version $\tilde{\mathbf{M}} = \mathbf{M} + \mathbf{P}$, where $\mathbf{P}$ is an arbitrary perturbation matrix. With this, PCUR computes the low rank approximation of $\mathbf{M}$ as $\mathbf{M} \approx \tilde{\mathbf{C}}\tilde{\mathbf{C}}^{\dagger}\tilde{\mathbf{M}}\tilde{\mathbf{R}}^{\dagger}\tilde{\mathbf{R}}$, where $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{R}}$ are column and row submatrices of $\tilde{\mathbf{M}}$, respectively, i.e., $\tilde{\mathbf{C}} = \tilde{\mathbf{M}}(:, J)$ and $\tilde{\mathbf{R}} = \tilde{\mathbf{M}}(I, :)$ for some index sets I, J. Therefore, the original PCUR projects a perturbed true matrix $\tilde{\mathbf{M}}$ onto the span of the columns of $\tilde{\mathbf{C}}$ and the rows of $\tilde{\mathbf{R}}$. We note that PCUR requires the **complete** knowledge of the noisy matrix, $\tilde{\mathbf{M}}$.

To evaluate the ability of PCUR to exploit the side information, we consider a variant of PCUR, called PCUR-p, where p indicates how many rows of $\hat{\mathbf{Q}}\mathbf{S}$ are provided as an input to PCUR (described in Table III). To provide a fair comparison, all variants of PCUR-p have access to the d true sampled columns, i.e., $\tilde{\mathbf{C}} \equiv \mathbf{A}$. PCUR-100 uses all rows of $\hat{\mathbf{Q}}\mathbf{S} \in \mathbb{R}^{100\times 100}$, i.e., $\tilde{\mathbf{R}} \equiv \hat{\mathbf{Q}}\mathbf{S}$, while PCUR-25 and PCUR-5 randomly sample 25 and 5 rows of $\hat{\mathbf{Q}}\mathbf{S}$, respectively. We chose 25 rows as it leads to the best performance for PCUR for several different values p. We emphasize again that, all variants of PCUR have access to a perturbed, yet complete, version of the ground truth, $\tilde{\mathbf{M}}$.

TABLE III
SUMMARY OF INFORMATION PROVIDED TO QPMA AND PCUR

Algorithm	Column space	Full matrix	Row space
PCUR-100	$\mathbf{A}$	$\tilde{\mathbf{M}}$	$\hat{\mathbf{Q}}\mathbf{S}$
PCUR-25	$\mathbf{A}$	$\tilde{\mathbf{M}}$	25 rows of $\hat{\mathbf{Q}}\mathbf{S}$
PCUR-5	$\mathbf{A}$	$\tilde{\mathbf{M}}$	5 rows of $\hat{\mathbf{Q}}\mathbf{S}$
QPMA	$\mathbf{A}$	-	$\hat{\mathbf{V}}_{QS}^{\mathsf{T}}$

A. Synthetic Data

Data generation. We generate the data as follows: The entries of the polynomial coefficient matrix, $\mathbf{Q} \in \mathbb{R}^{n\times l}$ are drawn i.i.d. from $\mathcal{N}(0, 1)$. We generate the reaction coordinate values, $\mathbf{s} \in \mathbb{R}^m$ as $[1 + 0.01 * [m]]^{\mathsf{T}} = [1.01, 1.02, \cdots, 1 + 0.01m]^{\mathsf{T}}$ and subsequently, the side information matrix $\mathbf{S}$ as in (3). For all experiments, we set $n = 100$ and $m = 100$. Next, in order to simultaneously control the "noise level" and the rank of the true matrix, $\mathbf{M}$, we generate the perturbation matrix, $\mathbf{E}$ as follows. Recall that $\mathbf{QS} \overset{\text{SVD}}{=} \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^{\mathsf{T}}$. We set $\mathbf{E} = \tilde{\mathbf{U}}_{[k]}\tilde{\mathbf{R}}_1\tilde{\mathbf{V}}_{[k]}^{\mathsf{T}}$, where $\tilde{\mathbf{U}}_{[k]} \in \mathbb{R}^{n\times k}$ the matrix of the first k columns of $\tilde{\mathbf{U}}$ and similarly for $\tilde{\mathbf{V}}_{[k]}$. The entries of $\tilde{\mathbf{R}}_1 \in \mathbb{R}^{k\times k}$ are drawn i.i.d. from $\mathcal{N}(0, \sigma_1^2)$, $\sigma_1^2 = 0.0001$. Recall that PCUR considers a perturbed version of $\mathbf{M}$, $\tilde{\mathbf{M}} = \mathbf{M} + \mathbf{P}$, where $\mathbf{P}$ is an arbitrary perturbation matrix. We let $\mathbf{P}$ be generated from the same distribution as $\mathbf{E}$, that is $\mathbf{P} = \tilde{\mathbf{U}}_{[k]}\tilde{\mathbf{R}}_2\tilde{\mathbf{V}}_{[k]}^{\mathsf{T}}$, where $\tilde{\mathbf{R}}_2 \in \mathbb{R}^{k\times k}$ is drawn i.i.d from $\mathcal{N}(0, \sigma_2^2)$, and $\sigma_2^2 = 0.0001$.

Varying d. We first investigate the performance of all algorithms as a function of the number of samples, governed by d. In the first experiment, we consider two values of the true rank $k = \{30, 50\}$, two possible polynomial degrees $l = \{3, 5\}$, and a noise standard deviation of $\sigma = 10^{-4}$. We note that the overall noise level is much higher since our main result and numerical results are derived with respect to the Frobenius norm.

We implement QPMA (Algorithm 1) with fixed step-size $\eta = 0.01$, and the maximum number of iterations $T = 1500$. We implement all variants of CUR+ with default parameters and we set $T = 1500$ to provide a fair comparison with QPMA. We plot the normalized mean square error (NMSE), $\|\mathbf{M} - \hat{\mathbf{M}}\|_F / \|\mathbf{M}\|_F$ for all algorithms in Fig. 3.

We notice from Fig. 3 that for both values of k, despite observing much fewer samples than CUR-H, the performance of QPMA is comparable to that of CUR-H. Moreover, QPMA outperforms all PCUR algorithms, despite the fact that PCUR has access to a perturbed version of the complete ground truth. This is possibly due to the fact that unlike QPMA, PCUR does not effectively leverage the quasi-polynomial side information. Furthermore, in the low-sample regime, i.e., $d \leq 10$, we observe that QPMA, CUR-H and all PCUR are almost two/three orders of magnitude better than CUR-S and CUR-L, which suggests that QPMA effectively exploits the quasi-polynomial side information. Although PCUR outperforms QPMA in the

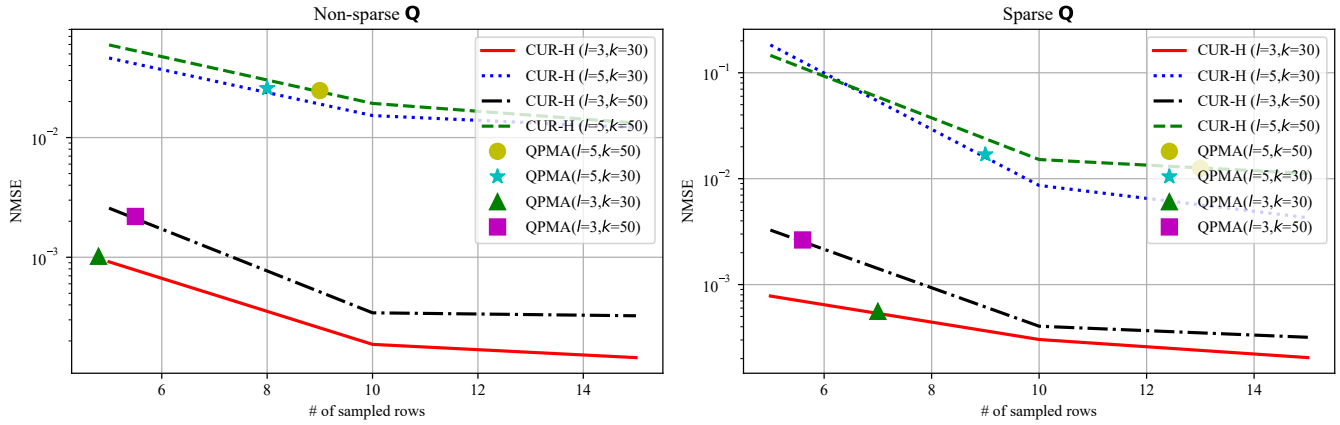

Fig. 5. Characterizing the informativeness of the quasi-polynomial side information through numerical performance of CUR+ when $d = 5$ and $r = 5$

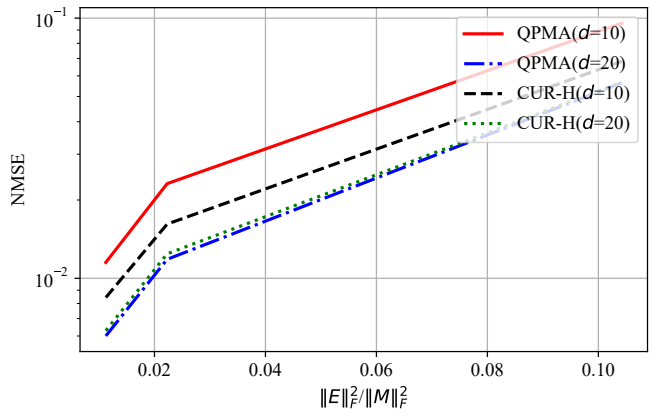

Fig. 6. The sensitivity of QPMA to the perturbation of $\mathbf{M}$ when $d = 10$ and $d = 20$. The parameters are set by $r = 5$, $l = 5$ and $k = 100$.

very low-sample regime, we observe that the performance of PCUR plateaus as the number of observed columns increases and we believe that this is due the fact that the reconstruction error in $\hat{\mathbf{Q}}\mathbf{S}$ does not significantly improve and thus using all rows of $\hat{\mathbf{Q}}\mathbf{S}$ is not beneficial. This again captures the assumption that the true matrix is only quasi-polynomial and not truly polynomial. QPMA (Algorithm 1), on the other hand, improves its performance as it observes more true columns and effectively optimizes for $\mathbf{Z}$ in (11).

Finally, we notice that as the number of columns, d, approaches the true rank, k, the performance of all algorithms do not significantly improve. This is in agreement with Theorem 1 since in this regime, the reconstruction error is dominated by the presence of noise, $\mathbf{E}$. Since the performance of PCUR is roughly similar to that of CUR+, we will only consider comparisons of QPMA with CUR+ for subsequent synthetic data experiments.

Varying k. Next, we analyze the effect of varying the true rank, k. We generate the data exactly as done in the first experiment with $l = 5$, $d = 10$. We set the target rank $r = 5$ for all algorithms. The results are provided in Fig. 4. Notice that when $k = l$, both QPMA and CUR-H are able to obtain near-perfect estimates of the true matrix with just $d = 2r = 10$ columns. As expected, the error increases with increasing k (since the number of observed columns and the polynomial degree is fixed), but saturates after $k \approx 3r$. Again, this observation is consistent with Theorem 1, as the error is dominated by the noise term, i.e., terms related to SNR rather than the approximation error $\|\mathbf{M} - \mathbf{M}_r\|_2^2$.

Sensitivity to Noise. We next investigate the sensitivity of QPMA to additional noise. We generate the data as done previously with $k = 100$, $d = \{10, 20\}$, and vary the noise standard deviation, $\sigma = \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. We provide the results in Fig. 6. As expected, the performance of both QPMA and CUR-H degrades as the noise increases. Furthermore, observe that for $d = 10$, CUR-H is more robust to noise. We believe that this is because in the regime of low d, the unrecoverable energy term of Theorem 1 dominates, while CUR-H has a lower effective bound due to observing significantly more samples than QPMA. For $d = 20$, we notice that QPMA is at least as robust as CUR-H and this is in accordnace with Theorem 1 as well, as in this regime error is dominated by the imperfect side-information (large $\mathbf{E}$) terms.

Quantifying row equivalence of side-information. We next attempt to answer the following question: *how much (row space) information is being captured by the quasi-polynomial side information*[7]. To this end, for both QPMA and CUR-H, we fix the number of observed columns to $d = 5$ and numerically compute the number of *rows* required for CUR-H to attain the same (fixed) numerical error as that of QPMA. Additionally, we consider two cases: the polynomial coefficient matrix, $\mathbf{Q}$ is **dense** (generated as in the previous experiment), and $\mathbf{Q}$ is **sparse** (generated by randomly puncturing 30% of the entries). The rest of the data is generated as before, with parameters $l = \{3, 5\}$ and $k = \{30, 50\}$. The results for these experiments are shown in Fig. 5. First, consider the dense $\mathbf{Q}$ case: for both values of k, observe that when $l = 5$, CUR-H requires roughly $8 - 9$ rows to match the error attained by Algorithm 1 and similarly when $l = 3$, CUR-H requires $5 - 6$ rows to match the error of QPMA. Thus, for dense $\mathbf{Q}$, CUR-H requires $\approx 2l$ rows to match the numerical

[7]For brevity, we only compare with CUR-H since the performance of QPMA is comparable to CUR-H across various parameter regimes.

performance of the proposed method. For sparse $\mathbf{Q}$, the effect is more pronounced, and CUR-H requires roughly $2l-3l$ rows to match the performance of QPMA. These observations also consistent with Theorem 1. With all other parameters fixed, making $\mathbf{Q}$ sparse, reducing k, and reducing l each have the effect of increasing δ and hence decreasing the (bound on) the error attained by QPMA.

B. Real data

We evaluate QPMA on the real Hessian eigenvalues matrix of a chemical system provided in [16]. In particular, we consider the CF_3CH_3 reaction system. For this system, the true matrix, $\mathbf{M} \in \mathbb{R}^{24\times 52}$, we observed that there is a good singular value separation at $r = 5$ and more specifically, $\sigma_1(\mathbf{M}) = 4.857$, $\sigma_6(\mathbf{M}) = 5.401 \times 10^{-3}$ and the matrix is full rank, i.e., $k = 24$. Informed by the singular value gap, we chose the target rank $r = 5$. Based on the methodology proposed in [16], we selected $l = 5$ and $s = [1 + 0.01 * [m]]$. For more detailed data description, see [16].

To simulate the setting of column-sampling only and limit the access to true rows, in the implementation of CUR-H, we provide d estimated rows (via QPMA) and $24 - d$ true rows. For PCUR-p, we set $p = 10$ as this resulted in the best performance for the CF_3CH_3 system. We also note that PCUR-25 is no longer implementable due to the dimension of the true matrix. To create a perturbed version of the true matrix $\tilde{\mathbf{M}}$, we generate an arbitrary perturbation matrix $\mathbf{P} \in \mathbb{R}^{24\times 52}$ where each entry of P is drawn i.i.d. from $\mathcal{N}(0, \sigma^2)$, with $\sigma = 0.07$. We choose the above value of σ so as to ensure $\|\mathbf{P}\|_F \doteq \|\mathbf{M} - \hat{\mathbf{Q}}\mathbf{S}\|_F$. QPMA is implemented with $\eta = 0.01$ and $T = 1500$. We present the results in Fig. 7. First, we observe that QPMA outperforms PCUR-10 in almost every regime ($d \geq 9$). The NMSE of PCUR-10 increases along with the number sampled columns d, while QPMA improves. We believe this is due to the fact that the row-space estimate for PCUR-10 does not improve with increasing d whereas QPMA optimizes $\mathbf{Z}$ more efficiently with increasing d. Additionally, we note that in the low-sample regime ($d \leq 13$), QPMA also outperforms CUR-H indicating that the side information is effectively being exploited by QPMA, whereas in the large sample regime, CUR-H tends to perform better than QPMA. However, we emphasize that (a) sampling more columns is often prohibitively more expensive in practice, and (b) CUR-H cannot be implemented in reality, since in general, one does not have access to the row information. Thus, we see that the proposed algorithm does in fact work well for the application that motivated the our algorithm. QPMA provides a tool by which the computation of key quantities for VTST can be reduced while offering good approximation performance. Our theoretical analysis provides strategies by which to understand VTST from a signal processing perspective.

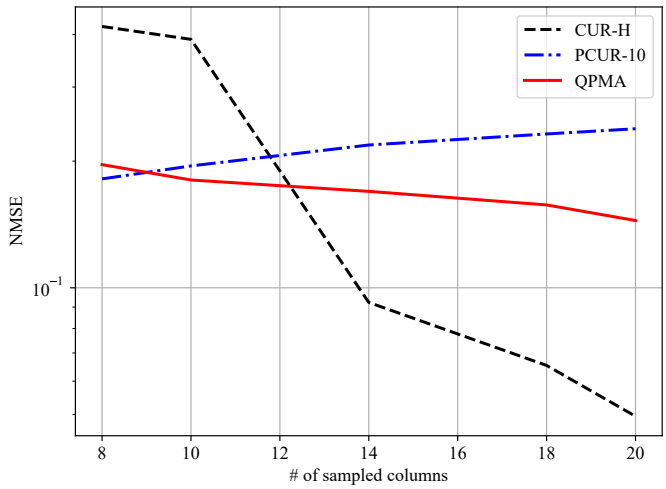

Fig. 7. The NMSE versus the number of sampled columns d for CF3CH3 chemical system with $l = 5$, $r = 5$ and $k = 24$.

VI. Conclusions

In this paper, we formulated a novel matrix approximation problem wherein we observe are a few arbitrary columns of a high-rank matrix. In order to make the problem tractable, and inspired by problems in quantum chemistry, we imposed a quasi-polynomial structural information. We designed and analyzed an algorithm dubbed Quasi-Polynomial Matrix Approximation (QPMA) to solve the above problem and derived theoretical guarantees. Our main guarantees show that the results are only slightly worse than state-of-the-art results in matrix approximation, albeit this work considers a significantly harder problem. Finally, we also provided several numerical experiments that validate our main guarantees. Specifically, we showed that (i) in the low-sample regime, the proposed method is roughly two to three orders of magnitude better than CUR+ [19]; (ii) the proposed algorithm outperforms PCUR [22] by effectively exploiting the side information, although both algorithms share the same column space and row space; (iii) in general, the polynomial structural information with degree l is roughly equivalent to observing $2l - 3l$ rows of the original matrix; and (iv) choosing the appropriate target rank is critical due to the sensitivity of the matrix approximation strategies to rank mismatch. Via simulation, it is shown that the error saturates after $k \approx 3r$. Finally, we show that our proposed methods work for the motivating quantum chemistry problem. We propose to characterize the classes of side information $\mathbf{S}$ that our approach can handle in future work.

Acknowledgements

We thank the anonymous reviewer who enabled a key improvement in Lemma 3. This work is funded in part by one or more of the following grants: NSF CCF-1817200, ARO W911NF1910269, DOE DE-SC0021417, Swedish Research Council 2018-04359, NSF CCF-2008927, NSF CCF-2200221, ONR 503400-78050, ONR N00014-15-1-2550 and USC+Amazon Trust AI center.

Appendix

The Appendix is organized as follows. First, we prove Theorem 1 in Appendix A, then we state and prove the key supporting Lemmas in Appendix B. Finally, in Appendix C we provide additional numerical experiments to address

the concerns of the anonymous reviewers from our previous submission.

APPENDIX A
PROOF OF THEOREM 1

Before proving Theorem 1, we first prove Lemmas 2 and 3. Throughout the proof, we invoke the following norm property. For a matrix $\mathbf{A}$, the $\|\cdot\|_2$ norm is given as,

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}_2\|=1} \|\mathbf{A}\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sigma_1(\mathbf{A}).$$

For a projection matrix, $\mathbf{P}_{\mathbf{U}_A}$ it is easy to see that $\sigma_1\left(\mathbf{P}_{\mathbf{U}_A}\right) = 1$.

Finally, we use Lemma 4 and Lemma 5 to prove Lemma 1 as follows.

$$\begin{aligned}
&\left\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2 \\
&\overset{(a)}{\leq} 2\left\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\right\|_2^2 + 2\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2 \\
&\overset{(b)}{\leq} 2\sigma_{r+1}^2(\mathbf{M})\left(3 + \frac{n}{d}\right) \\
&\qquad + 32\sigma_1^2(\mathbf{M})\left\|\mathbf{E}\right\|_F^2\left(\frac{1}{\delta_1^2} + \frac{\left\|(\mathbf{S}\mathbf{\Psi})^\dagger\mathbf{S}\right\|_F^2}{\delta_2^2}\right) \qquad (19)
\end{aligned}$$

The inequalities (a) is due to the fact that $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \geq 0$. (b) follows from Lemma 4 and Lemma 5. This concludes the proof of Lemma 1.

A. Proof of Lemma 3

Lemma 3 ensures the strong convexity of the objective function $f(\mathbf{Z})$ in (12) by restricting the curvature of the column sampling operator $\mathbf{\Psi}$ [18], [41]. Recall that $\mathbf{\Psi}$ is consisted of d number of randomly chosen columns in $\mathbf{M}$. We first provide an additional necessary theorem from [35] which describes the large-deviation behavior of specific types of matrix random variables.

Theorem 2. *[35] Let $\mathcal{X}$ be a finite set of positive semi definite matrices of dimension k. If there exists a constant $B < \infty$ such that*

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

and, if we sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_p\}$ uniformly at random from $\mathcal{X}$ without replacement, with

$$\begin{aligned}
\mu_{max} &:= p\lambda_{max}(\mathbb{E}[\mathbf{X}_1]) \\
\mu_{min} &:= p\lambda_{min}(\mathbb{E}[\mathbf{X}_1]).
\end{aligned}$$

Then we have that,

$$\begin{aligned}
&P\left[\lambda_{max}\left(\sum_{i=1}^{p}\mathbf{X}_i\right) \geq (1+\rho)\mu_{max}\right] \\
&\leq k \cdot exp\frac{-\mu_{max}}{B}[(1+\rho)\ln(1+\rho) - \rho] \text{ for } \rho \in [0,1) \\
&P\left[\lambda_{min}\left(\sum_{i=1}^{p}\mathbf{X}_i\right) \leq (1-\rho)\mu_{min}\right] \\
&\leq k \cdot exp\frac{-\mu_{min}}{B}[(1-\rho)\ln(1-\rho) + \rho] \text{ for } \rho \in [0,1)
\end{aligned}$$

$\boxtimes$

Recall that Lemma 3 provides a bound on the value of α in our definition of strong convexity in Definition 2. A way to prove a function is strictly convex is to show the Hessian of a function is strictly positive definite [34] everywhere. We can show the Hessian matrix is positive definite by bounding the smallest eigenvalue of the Hessian matrix as a positive value.

Observe that $f(\mathbf{Z})$ can be expressed as

$$f(\mathbf{Z}) = \left\|\mathbf{A} - \mathbf{U}_A\mathbf{Z}\hat{\mathbf{V}}_{QS}^\mathsf{T}\mathbf{\Psi}\right\|_F^2.$$

In order to show that the smallest value of the Hessian of $f(\mathbf{Z})$ is bounded away from zero, notice that $f(\mathbf{Z})$ can be expressed as

$$\begin{aligned}
f(\mathbf{Z}) &= \left\|\mathbf{A} - \left(\mathbf{U}_A\mathbf{Z}\hat{\mathbf{V}}_{QS}^\mathsf{T}\mathbf{\Psi}\right)\right\|_F^2 \\
&= \left\|\sum_{j\in\mathcal{C}}\mathbf{m}_j - \left(\mathbf{U}_A\mathbf{Z}\hat{\mathbf{V}}_{QS}^\mathsf{T}\right)_j\right\|_F^2 \\
&\overset{(a)}{=} \sum_{j\in\mathcal{C}}\left\|\mathbf{m}_j - \left(\mathbf{U}_A\mathbf{Z}\hat{\mathbf{V}}_{QS}^\mathsf{T}\right)_j\right\|_F^2,
\end{aligned}$$

where (a) follows from the definition of the Frobenius norm; and using $\mathbf{m}_j$ to denote the j-th column of $\mathbf{M}$. Similarly, letting $\hat{\mathbf{m}}_j$ indicate the j-th column of $\hat{\mathbf{M}}$, we have,

$$\begin{aligned}
f(\mathbf{Z}) &= \sum_{j\in\mathcal{C}}\|\mathbf{m}_j - \hat{\mathbf{m}}_j\|_F^2, \\
&= \sum_{j\in\mathcal{C}} f_j(\mathbf{Z}), \qquad (20)
\end{aligned}$$

where $f_j(\mathbf{Z}) = \|\mathbf{m}_j - \hat{\mathbf{m}}_j\|_F^2 = \left\|\mathbf{m}_j - \left(\mathbf{U}_A\mathbf{Z}\hat{\mathbf{V}}_{QS}^\mathsf{T}\right)_j\right\|_F^2$. Now, taking the second-order derivative with respect to each element z_{th} and z_{pq} for $t, p \in [r]$ and $h, q \in [r]$, we have

$$\frac{\partial^2 f(\mathbf{Z})}{\partial z_{th}\partial z_{pq}} = \sum_{j\in\mathcal{C}}\frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{th}\partial z_{pq}}.$$

We define

$$\mathbf{H} := \frac{\partial^2 f(\mathbf{Z})}{\partial z_{th}\partial z_{pq}} \text{ and } \mathbf{H}_j := \frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{th}\partial z_{pq}}, \qquad (21)$$

therefore, by linearity of the derivative it is easy to see that

$$\mathbf{H} = \sum_{j\in\mathcal{C}}\mathbf{H}_j. \qquad (22)$$

Furthermore, the first-order derivative of $\frac{\partial f_j(\mathbf{Z})}{\partial z_{th}}$ with respect to z_{th} for $t \in [r]$ and $h \in [r]$, is given by

$$\begin{aligned}
\frac{\partial f_j(\mathbf{Z})}{\partial z_{th}} = -2\sum_{i\in[n]}\left(m_{ij} - \sum_{t\in[r]}\sum_{h\in[r]}u_{A,it}z_{th}\hat{v}_{QS,hj}\right) \\
\cdot u_{A,it}\hat{v}_{QS,hj}.
\end{aligned}$$

And subsequently, the second-derivative of $f_j(\mathbf{Z})$ with respect to z_{th}, z_{pq} for $t, p \in [r]$ and $h, q \in [r]$ can be expressed as

$$\frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{th}\partial z_{pq}} = 2\sum_{i\in[n]}u_{A,it}\hat{v}_{QS,hj}u_{A,ip}\hat{v}_{QS,qj}.$$

We let the second-order derivative of the (i_1, j_1)th and (i_2, j_2) entry of $\mathbf{Z}$ be the $(r(i_1-1)+j_1, r(i_2-1)+j_2)$ entry of the Hessian matrix of $f_j(\mathbf{Z})$. Then, the Hessian $\mathbf{H}_j$ of $f_j(\mathbf{Z})$ is

$$\mathbf{H}_j \in \mathbb{R}^{r^2\times r^2} = \begin{bmatrix} \frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{11}\partial z_{11}} & \frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{11}\partial z_{12}} & \cdots & \frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{11}\partial z_{rr}} \\ \frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{12}\partial z_{11}} & \frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{12}\partial z_{12}} & \cdots & \frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{12}\partial z_{rr}} \\ \vdots & \vdots & \vdots & \ddots \\ \frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{rr}\partial z_{11}} & \frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{rr}\partial z_{12}} & \cdots & \frac{\partial^2 f_j(\mathbf{Z})}{\partial z_{rr}\partial z_{rr}}. \end{bmatrix}$$

Combining all these, the Hessian matrix of $f_j(\mathbf{Z})$ can be succinctly expressed as,

$$\mathbf{H}_j = 2\sum_{i\in[n]} \left[\text{vec}(\mathbf{u}_{A,i}^\mathsf{T}\hat{\mathbf{v}}_{QS,j})\right]\left[\text{vec}(\mathbf{u}_{A,i}^\mathsf{T}\hat{\mathbf{v}}_{QS,j})\right]^\mathsf{T},$$

where $\mathbf{u}_{A,i}$, for $i \in [n]$, is the i-th row of $\mathbf{U}_A$ and $\hat{\mathbf{v}}_{QS,j}$ for $j \in [m]$, is the j-th row of $\hat{\mathbf{V}}_{QS}$. We now bound the smallest eigenvalue of $\mathbf{H}$ using Theorem 2. In order to invoke Theorem 2, we consider $\mathbf{H} = \sum_j \mathbf{H}_j$ for $j \in \mathcal{C}$, where $\mathbf{H}_j \equiv \mathbf{X}_i$ and since we are sampling $|\mathcal{C}| = d$ coordinates sampled without replacement, as done in [35, Proof of Lemma 3.4] $\mathbf{H}$ is equivalently represented as a sum of d random matrices. Next, in order to apply Theorem 2, we derive an upper bound on the largest eigenvalue of $\mathbf{H}_j$ as

$$\begin{aligned}
&\max_j \lambda_{\max}\left(\sum_{i\in[n]} 2\left[\text{vec}(\mathbf{u}_{A,i}^\mathsf{T}\hat{\mathbf{v}}_{QS,j})\right]\left[\text{vec}(\mathbf{u}_{A,i}^\mathsf{T}\hat{\mathbf{v}}_{QS,j})\right]^\mathsf{T}\right) \\
&\overset{(a)}{\leq} 2\max_j \sum_{i\in[n]} \lambda_{\max}\left(\left[\text{vec}(\mathbf{u}_{A,i}^\mathsf{T}\hat{\mathbf{v}}_{QS,j})\right]\left[\text{vec}(\mathbf{u}_{A,i}^\mathsf{T}\hat{\mathbf{v}}_{QS,j})\right]^\mathsf{T}\right) \\
&\overset{(b)}{\leq} \max_j \sum_{i\in[n]} \left\|\text{vec}\left(\mathbf{u}_{A,i}^\mathsf{T}\hat{\mathbf{v}}_{QS,j}\right)\right\|_2^2 \\
&\overset{(c)}{\leq} 2\max_j \sum_{i\in[n]} \left\|\mathbf{u}_{A,i}^\mathsf{T}\hat{\mathbf{v}}_{QS,j}\right\|_F^2 \\
&\overset{(d)}{\leq} 2\sum_{i\in[n]} \|\mathbf{u}_{A,i}\|_F^2 \max_j \|\hat{\mathbf{v}}_{QS,j}\|_F^2 \overset{(e)}{\leq} \frac{2\hat{\mu}^2 r^2}{m} := B,
\end{aligned}$$

where (a) follows from Weyl's inequality [42], (b) follows since the argument of $\lambda_{\max}$ is the outer product of two vectors; (c) follows from norm inequalities; (d) is from the Cauchy-Schwarz inequality and (e) is due to the definition of the incoherence of a matrix defined in Definition 1.

Next, we have

$$\begin{aligned}
\mathbb{E}[\mathbf{H}_1^{'}] &= \frac{2}{m}\sum_{j=1}^{m}\sum_{i\in[n]} \left[\text{vec}(\mathbf{u}_{A,i}^\mathsf{T}\hat{\mathbf{v}}_{QS,j})\right]\left[\text{vec}(\mathbf{u}_{A,i}^\mathsf{T}\hat{\mathbf{v}}_{QS,j})\right]^\mathsf{T} \\
&= \frac{2}{m}\left(\left(\mathbf{U}_A\otimes\hat{\mathbf{V}}_{QS}\right)^\mathsf{T}\times\left(\mathbf{U}_A\otimes\hat{\mathbf{V}}_{QS}\right)\right) = \frac{2}{m}\mathbf{I}_{r^2}
\end{aligned}$$

where $\otimes$ denotes the Kronecker product and the equality follows from the orthogonality of the columns. Thus, $\lambda_{\min}(\mathbb{E}[\mathbf{H}_1]) = 2/m$. Finally, with $B = \frac{2\hat{\mu}^2 r^2}{m}$, $\mu_{\min} = \frac{2d}{m}$ and $\rho = \frac{1}{2}$, we have,

$$P\left\{\lambda_{\min}(\mathbf{H}) \leq \frac{d}{2m}\right\} \leq \frac{r^2}{6} e^{\frac{-d}{\hat{\mu}^2 r^2}} = e^{\frac{-d}{\hat{\mu}^2 r^2} + \frac{\ln r}{3}}.$$

This expression can be algebraically manipulated, such that with a probability $1 - r^{-c_1}$, where $c_1 > 0$ is a constant, and if $d \geq c_2\hat{\mu}^2 r^2 \ln r$ for a constant $c_2 > 0$, we have,

$$\lambda_{\min}(\mathbf{H}) \geq \frac{d}{2m}.$$

B. Proof of Lemma 2

In Lemma 2, we bound the error between our projected true matrix and our final estimate of the reconstructed matrix as follows. We note that $\hat{\mathbf{M}} = \mathbf{U}_A\hat{\mathbf{Z}}\hat{\mathbf{V}}_{QS}^\mathsf{T}$ and hence $\hat{\mathbf{Z}} = \mathbf{U}_A^\mathsf{T}\hat{\mathbf{M}}\hat{\mathbf{V}}_{QS}$. Also recall that $\mathbf{Z} = \mathbf{U}_A^\mathsf{T}\mathbf{M}\hat{\mathbf{V}}_{QS}$.

$$\begin{aligned}
&\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}} - \mathbf{U}_A\hat{\mathbf{Z}}\hat{\mathbf{V}}_{QS}^\mathsf{T}\right\|_2^2 \\
&= \left\|\mathbf{U}_A\mathbf{U}_A^\mathsf{T}\mathbf{M}\hat{\mathbf{V}}_{QS}\hat{\mathbf{V}}_{QS}^\mathsf{T} - \mathbf{U}_A\hat{\mathbf{Z}}\hat{\mathbf{V}}_{QS}^\mathsf{T}\right\|_2^2 \\
&\overset{(a)}{=} \|\mathbf{U}_A\mathbf{Z}\hat{\mathbf{V}}_{QS} - \mathbf{U}_A\hat{\mathbf{Z}}\hat{\mathbf{V}}_{QS}^\mathsf{T}\|_2^2 \\
&\overset{(b)}{\leq} \|\mathbf{U}_A\|_2^2\left\|\mathbf{Z}-\hat{\mathbf{Z}}\right\|_2^2\left\|\hat{\mathbf{V}}_{QS}\right\|_2^2 \\
&\overset{(c)}{=} \left\|\mathbf{Z}-\hat{\mathbf{Z}}\right\|_2^2, \qquad (23)
\end{aligned}$$

where (a) is obtained from the definition of $\mathbf{Z}$, and (b) from matrix norm inequalities. As $\mathbf{U}_A$ and $\hat{\mathbf{V}}_{QS}$ are unitary, their 2-norms are unity (c).

Finally, $\left\|\mathbf{Z}-\hat{\mathbf{Z}}\right\|_2^2$ is bounded as follows. Recall in Lemma 3, we established the lower bound on convergence rate α that ensures the strong convexity of $f(\mathbf{Z})$. This result, in turn, let us establish the error bound for $\left\|\mathbf{Z}-\hat{\mathbf{Z}}\right\|_2^2$, which provides the bound for sample complexity. We have

$$\begin{aligned}
\frac{\alpha}{2}\left\|\mathbf{Z}-\hat{\mathbf{Z}}\right\|_2^2 &\overset{(a)}{\leq} \left\|\mathbf{M}\mathbf{\Psi} - \mathbf{U}_A\mathbf{Z}\hat{\mathbf{V}}_{QS}^\mathsf{T}\mathbf{\Psi}\right\|_F^2 - \left\|\mathbf{M}\mathbf{\Psi} - \mathbf{U}_A\hat{\mathbf{Z}}\hat{\mathbf{V}}_{QS}^\mathsf{T}\mathbf{\Psi}\right\|_F^2 \\
&\leq \left\|\mathbf{M}\mathbf{\Psi} - \mathbf{U}_A\mathbf{Z}\hat{\mathbf{V}}_{QS}^\mathsf{T}\mathbf{\Psi}\right\|_F^2 \\
&\overset{(b)}{=} \left\|\mathbf{M}\mathbf{\Psi} - \mathbf{U}_A\mathbf{U}_A^\mathsf{T}\mathbf{M}\hat{\mathbf{V}}_{QS}\hat{\mathbf{V}}_{QS}^\mathsf{T}\mathbf{\Psi}\right\|_F^2 \\
&= \left\|\mathbf{M}\mathbf{\Psi} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\mathbf{\Psi}\right\|_F^2 \\
&\leq \left\|\mathbf{M} - \mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_F^2 \\
&\overset{(c)}{\leq} 2\sigma_{r+1}^2(\mathbf{M})\left(3+\frac{n}{d}\right) \\
&\quad + 32\sigma_1^2(\mathbf{M})\|\mathbf{E}\|_F^2\left(\frac{1}{\delta_1^2} + \frac{\left\|(\mathbf{S}\mathbf{\Psi})^\dagger\mathbf{S}\right\|_F^2}{\delta_2^2}\right) \qquad (24)
\end{aligned}$$

where (a) is from $f(\mathbf{Z}) - f(\hat{\mathbf{Z}}) \geq \frac{\alpha}{2}\left\|\mathbf{Z}-\hat{\mathbf{Z}}\right\|_2^2$ in Definition 2. Since Gradient Descent reaches a stationary point, it follows that $\bigtriangledown f(\hat{\mathbf{Z}}) = 0$. And (b) is from our definition of $\mathbf{Z}$, $\mathbf{Z} = \mathbf{U}_A^\mathsf{T}\mathbf{M}\hat{\mathbf{V}}_{QS}$. Finally, (c) follows from Lemma 1.

$$\begin{aligned}
\Delta \equiv\ & 2\sigma_{r+1}^2(\mathbf{M})\left(3+\frac{n}{d}\right) \\
&+ 32\sigma_1^2(\mathbf{M})\|\mathbf{E}\|_F^2\left(\frac{1}{\delta_1^2} + \frac{\left\|(\mathbf{S}\mathbf{\Psi})^\dagger\mathbf{S}\right\|_F^2}{\delta_2^2}\right) \qquad (25)
\end{aligned}$$

Then, we have,

$$\left\|\mathbf{Z}-\hat{\mathbf{Z}}\right\|_F^2 \leq \frac{2\Delta}{\alpha}.$$

By plugging $\left\|\mathbf{Z}-\hat{\mathbf{Z}}\right\|_F^2 \leq \frac{2\Delta}{\alpha}$ into (23) and replacing Δ with (25), we obtain the bound of $\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}-\mathbf{U}_A\hat{\mathbf{Z}}\hat{\mathbf{V}}_{QS}^{\mathsf{T}}\right\|_2^2$ in Lemma 2 as

$$\begin{aligned}&\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}-\mathbf{U}_A\hat{\mathbf{Z}}\hat{\mathbf{V}}_{QS}^{\mathsf{T}}\|_2^2\\ &\leq \frac{4m}{d}\left(2\sigma_{r+1}^2(\mathbf{M})\left(3+\frac{n}{d}\right)\right.\\ &\qquad\left.+\,32\sigma_1^2(\mathbf{M})\|\mathbf{E}\|_F^2\left(\frac{1}{\delta_1^2}+\frac{\left\|(\mathbf{S}\boldsymbol{\Psi})^\dagger\mathbf{S}\right\|_F^2}{\delta_2^2}\right)\right). \quad (26)\end{aligned}$$

With $\alpha \geq \frac{d}{2m}$, We obtain Lemma 2 provided that

$$d \geq c_1\hat{\mu}^2r^2(\ln r),$$

and $c_1 \geq 0$.

Appendix B
Proof of auxiliary Lemmas

In this section, we first provide the proof of Lemma 1 followed by the proof of Lemma 5 that is key to proving the main result. The proof is based on a careful application of Wedin's Theorem [36] and some linear algebra.

Proof of Lemma 1. We use Lemma 4 and Lemma 5 to prove Lemma 1. Lemma 4 was proved in [19].

Lemma 4. *[19] With probability $1-c_1r^{-10}$, and if $d \geq c_2\mu r\ln r$, for constants $c_1>0$ and $c_2>0$, we have*

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

Lemma 5. *Assume that there exists a $\delta_1>0$ that satisfies Assumption 1. Then, if $d \geq c\mu r\ln r$, we have,*

$$\begin{aligned}&\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}-\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\ &\leq 2\sigma_{r+1}^2(\mathbf{M})+16\sigma_1^2(\mathbf{M})\|\mathbf{E}\|_F^2\left(\frac{1}{\delta_1^2}+\frac{1}{\delta_s^2}\right),\end{aligned}$$

where $\delta_s := \frac{\delta_2}{\left\|(\mathbf{S}\boldsymbol{\Psi})^\dagger\mathbf{S}\right\|_F}$ and $c>0$ is a constant.

We prove Lemma 5 next.

Proof of Lemma 5. We first define some preliminaries that are required to prove Lemma 5. We use the following definition of Canonical angles as a distance measure between subspaces.

Definition 3 (Canonical angle between subspaces [36])**.** *Let $\mathbf{X}\in\mathbb{R}^{n\times k}$ and $\mathbf{Y}\in\mathbb{R}^{n\times k}$ be matrices, whose columns form orthonormal basis for column space of each. Let $\gamma_1\geq\cdots\geq\gamma_k$ be the singular values of $\mathbf{X}^{\mathsf{T}}\mathbf{Y}$. Then, the canonical angles between the column subspace of $\mathbf{X}$ and $\mathbf{Y}$ are defined as*

$$\theta_i = \cos^{-1}\gamma_i,\ i\in[k].$$

Next, we introduce Wedin's theorem [36] that is used to bound the distance between the subspaces of two matrices. For this part, consider $\mathbf{X}\in\mathbb{R}^{n\times m}$ with rank k and let $\tilde{\mathbf{X}}=\mathbf{X}+\mathbf{E}$ be a perturbation of $\mathbf{X}$. Denote the SVDs of $\mathbf{X},\tilde{\mathbf{X}}$ as

$$\mathbf{X} \overset{\text{SVD}}{=} \mathbf{U}_1\boldsymbol{\Sigma}_1\mathbf{V}_1^{\mathsf{T}}+\mathbf{U}_2\boldsymbol{\Sigma}_2\mathbf{V}_2^{\mathsf{T}}, \quad (27)$$

and

$$\tilde{\mathbf{X}} \overset{\text{SVD}}{=} \tilde{\mathbf{U}}_1\tilde{\boldsymbol{\Sigma}}_1\tilde{\mathbf{V}}_1^{\mathsf{T}}+\tilde{\mathbf{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\mathbf{V}}_2^{\mathsf{T}}. \quad (28)$$

where the singular values are not necessarily presented in a descending order. Then, Wedin's theorem says the following.

Theorem 3 (**Wedin's theorem** [36])**.** *Let $\boldsymbol{\Phi}$ denote the matrix of canonical angles between $\mathbf{U}_1$ and $\tilde{\mathbf{U}}_1$, and $\boldsymbol{\Theta}$ be the matrix of canonical angles between $\mathbf{V}_1$ and $\tilde{\mathbf{V}}_1$ in* (27) *and* (28) *respectively. And, if there is a $\delta>0$ such that*

$$\delta = \min\{\min_{1\leq i\leq k,k\leq j\leq m}|\sigma_i(\mathbf{X})-\sigma_j(\tilde{\mathbf{X}})|,\min_{1\leq i\leq k}\sigma_i(\mathbf{X})\}$$

then

$$\sqrt{\|\sin\boldsymbol{\Theta}\|_F^2+\|\sin\boldsymbol{\Phi}\|_F^2} \leq \frac{\sqrt{\|\mathbf{R}_1\|_F^2+\|\mathbf{S}_1\|_F^2}}{\delta},$$

where $\mathbf{R}_1$, "residual between the column spaces" is

$$\mathbf{R}_1 = \mathbf{X}\tilde{\mathbf{V}}_1-\tilde{\mathbf{U}}_1\tilde{\boldsymbol{\Sigma}}_1$$

and $\mathbf{S}_1$, the "residual between the row spaces" is

$$\mathbf{S}_1 = \mathbf{X}^{\mathsf{T}}\tilde{\mathbf{U}}_1-\tilde{\mathbf{V}}_1\tilde{\boldsymbol{\Sigma}}_1.$$

⊠

We also use the following theorem that discusses the connection between canonical angles and projections.

Theorem 4. *(***The connection between canonical angles and projections** *[43]) Let $\mathbf{P}_{\mathbf{V}_1}$ and $\mathbf{P}_{\tilde{\mathbf{V}}_1}$ denote the orthogonal projections onto $\mathbf{V}_1$ and $\tilde{\mathbf{V}}_1$ respectively. Let $\boldsymbol{\Theta}$ be the matrix of canonical angles between $\mathbf{V}_1$ and $\tilde{\mathbf{V}}_1$. Define $\|\mathbf{P}_{\mathbf{V}_1}-\mathbf{P}_{\tilde{\mathbf{V}}_1}\|_F$. Then,*

$$\left\|\mathbf{P}_{\mathbf{V}_1}-\mathbf{P}_{\tilde{\mathbf{V}}_1}\right\|_F = \sqrt{2}\left\|\sin\boldsymbol{\Theta}\right\|_F. \quad (29)$$

⊠

Now, the proof of Lemma 5 follows from two applications of Theorem 3 and Theorem 4. For the first application, we invoke Theorem 3 with $\mathbf{X}\equiv\mathbf{Q}\mathbf{S}, \tilde{\mathbf{X}}\equiv\hat{\mathbf{Q}}\mathbf{S}$. Recall that $\mathbf{Q}$ indicates the true polynomial coefficient matrix from (2) and $\hat{\mathbf{Q}}$ is the estimate that is obtained from (9).Then, we have

$$\begin{aligned}\delta_2 &= \min\{\min_{1\leq i\leq r,r+1\leq j\leq m}|\sigma_i(\mathbf{Q}\mathbf{S})-\sigma_j(\hat{\mathbf{Q}}\mathbf{S})|,\min_{1\leq i\leq r}\sigma_i(\mathbf{Q}\mathbf{S})\}\\ &\overset{(a)}{=} \min_{1\leq i\leq r,r+1\leq j\leq m}|\sigma_i(\mathbf{Q}\mathbf{S})-\sigma_j(\hat{\mathbf{Q}}\mathbf{S})| \quad (30)\\ &= |\sigma_r(\mathbf{Q}\mathbf{S})-\sigma_{r+1}(\hat{\mathbf{Q}}\mathbf{S})| \quad (31)\end{aligned}$$

where (a) follows from using the fact that target rank $r\leq l$. In appendix, we investigate $\delta_2>0$ numerically. Next, we compute the residuals required for Wedin's theorem as follows

$$\mathbf{T} := \hat{\mathbf{Q}}\mathbf{S}\mathbf{V}_{QS}-\mathbf{U}_{QS}\boldsymbol{\Sigma}_{QS} \quad (32)$$

and

$$\mathbf{W} := (\hat{\mathbf{Q}}\mathbf{S})^{\mathsf{T}}\mathbf{U}_{QS} - \mathbf{V}_{QS}\mathbf{\Sigma}_{QS}. \tag{33}$$

Let

$$\mathbf{D} := \hat{\mathbf{Q}}\mathbf{S} - \mathbf{Q}\mathbf{S}. \tag{34}$$

Then, we invoke Theorem 4 with $\mathbf{V}_1 \equiv \mathbf{V}_{QS}$ and $\tilde{\mathbf{V}}_1 \equiv \hat{\mathbf{V}}_{QS}$ and let $\mathbf{\Theta_1}$ be the matrix of canonical angles between $\mathbf{V}_{QS}$ and $\hat{\mathbf{V}}_{QS}$ to obtain

$$\begin{aligned}
\left\|\mathbf{P}_{\mathbf{V}_{QS}} - \mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2 &= \left\|\mathbf{V}_{QS}\mathbf{V}_{QS}^{\mathsf{T}} - \hat{\mathbf{V}}_{QS}\hat{\mathbf{V}}_{QS}^{\mathsf{T}}\right\|_2^2 \\
&\overset{(a)}{\leq} \left\|\mathbf{V}_{QS}\mathbf{V}_{QS}^{\mathsf{T}} - \hat{\mathbf{V}}_{QS}\hat{\mathbf{V}}_{QS}^{\mathsf{T}}\right\|_F^2 \\
&= 2\|\sin\mathbf{\Theta_1}\|_F^2 \\
&\leq 2\left(\frac{\|\mathbf{T}\|_F^2 + \|\mathbf{W}\|_F^2}{\delta_2^2} - m\right) \\
&\leq 2\left(\frac{\|\mathbf{T}\|_F^2 + \|\mathbf{W}\|_F^2}{\delta_2^2}\right) \\
&\overset{(b)}{=} \frac{4\|\mathbf{D}\|_F^2}{\delta_2^2},
\end{aligned} \tag{35}$$

where (a) follows from using $\|\cdot\|_2^2 \leq \|\cdot\|_F^2$ and (b) is due to

$$\|\mathbf{T}\|_F^2 = \|\hat{\mathbf{Q}}\mathbf{S}\mathbf{V}_{QS} - \mathbf{U}_{QS}\mathbf{\Sigma}_{QS}\|_F^2 = \|\mathbf{D}\mathbf{V}_{QS}\|_F^2 = \|\mathbf{D}\|_F^2, \tag{36}$$

with a similar bound for $\mathbf{W}$. Now, we further bound $\|\mathbf{D}\|_F^2$ in (34). Given $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{\Psi}$, $\hat{\mathbf{Q}}$ is obtained by solving the unconditioned least-squares problem (8). This problem can be solved analytically. Since the rows of $\mathbf{S}\mathbf{\Psi}$ are independent, the least-squares approximation problem has the unique solution $\hat{\mathbf{Q}} = \mathbf{A}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\left(\mathbf{S}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\right)^{-1}$ [44, p.155]. Therefore, we have the following bound for $\|\mathbf{D}\|_F^2$,

$$\begin{aligned}
\|\mathbf{D}\|_F^2 &= \|\hat{\mathbf{Q}}\mathbf{S} - \mathbf{Q}\mathbf{S}\|_F^2 \\
&= \left\|\mathbf{A}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\left(\mathbf{S}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\right)^{-1}\mathbf{S} - \mathbf{Q}\mathbf{S}\right\|_F^2 \\
&= \left\|\mathbf{M}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\left(\mathbf{S}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\right)^{-1}\mathbf{S} - \mathbf{Q}\mathbf{S}\right\|_F^2 \\
&= \left\|(\mathbf{Q}\mathbf{S} + \mathbf{E})\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\left(\mathbf{S}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\right)^{-1}\mathbf{S} - \mathbf{Q}\mathbf{S}\right\|_F^2 \\
&= \|\mathbf{Q}\mathbf{S}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\left(\mathbf{S}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\right)^{-1}\mathbf{S} \\
&\qquad + \mathbf{E}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\left(\mathbf{S}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\right)^{-1}\mathbf{S} - \mathbf{Q}\mathbf{S}\|_F^2 \\
&= \left\|\mathbf{E}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\left(\mathbf{S}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\mathsf{T}}\right)^{-1}\mathbf{S}\right\|_F^2 \\
&= \left\|\mathbf{E}\mathbf{\Psi}\left(\mathbf{S}\mathbf{\Psi}\right)^{\dagger}\mathbf{S}\right\|_F^2 \\
&\overset{(a)}{\leq} \|\mathbf{E}\|_F^2\left\|\left(\mathbf{S}\mathbf{\Psi}\right)^{\dagger}\mathbf{S}\right\|_F^2,
\end{aligned} \tag{37}$$

where (a) is due to the matrix norm inequality.

Thus, using (37) and (35), we have,

$$\left\|\mathbf{P}_{\mathbf{V}_{QS}} - \mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2 \leq \frac{4\|\mathbf{E}\|_F^2}{\delta_2^2}\left\|\left(\mathbf{S}\mathbf{\Psi}\right)^{\dagger}\mathbf{S}\right\|_F^2. \tag{38}$$

Next, using a similar approach, we apply Theorems 3 and 4 with $\mathbf{X} \equiv \mathbf{M}$ and $\tilde{\mathbf{X}} \equiv \mathbf{Q}\mathbf{S}$.

By plugging (38) and (42) into (45), we have

$$\left\|\mathbf{P}_{\mathbf{V}_M} - \mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2 \leq 8\|\mathbf{E}\|_F^2\left(\frac{1}{\delta_1^2} + \frac{\left\|\left(\mathbf{S}\mathbf{\Psi}\right)^{\dagger}\mathbf{S}\right\|_F^2}{\delta_2^2}\right). \tag{39}$$

First, we demonstrate the minimum eigengap separation condition as follows. Let

$$\delta_1 = \min\{\min_{1\leq i\leq l, l+1\leq j\leq m}|\sigma_i(\mathbf{Q}\mathbf{S}) - \sigma_j(\mathbf{M})|, \min_{1\leq i\leq l}\sigma_i(\mathbf{Q}\mathbf{S})\}$$

Notice that since $\text{rank}(\mathbf{M}) = k \gg r$ and $\text{rank}(\mathbf{Q}\mathbf{S}) = l \geq r$, the first term above attains the minimum and thus $\delta_1 = \sigma_r(\mathbf{M} - \mathbf{E}) - \sigma_{r+1}(\mathbf{M})$. This is bounded away from zero owing to Assumption 1. We next compute the residuals as follows

$$\mathbf{R} := \mathbf{Q}\mathbf{S}\mathbf{V}_M - \mathbf{U}_M\mathbf{\Sigma}_M \tag{40}$$

and $\mathbf{S}$ is defined as the residual between the row space $\mathbf{V}_M$ and $\mathbf{V}_{QS}$ as

$$\mathbf{S} := (\mathbf{Q}\mathbf{S})^{\mathsf{T}}\mathbf{U}_M - \mathbf{V}_M\mathbf{\Sigma}_M. \tag{41}$$

Then,

$$\begin{aligned}
\left\|\mathbf{P}_{\mathbf{V}_M} - \mathbf{P}_{\mathbf{V}_{QS}}\right\|_2^2 &= \left\|\mathbf{V}_M\mathbf{V}_M^{\mathsf{T}} - \mathbf{V}_{QS}\mathbf{V}_{QS}^{\mathsf{T}}\right\|_2^2 \\
&\overset{(a)}{\leq} \left\|\mathbf{V}_M\mathbf{V}_M^{\mathsf{T}} - \mathbf{V}_{QS}\mathbf{V}_{QS}^{\mathsf{T}}\right\|_F^2 \\
&\overset{(b)}{=} 2\|\sin\mathbf{\Theta}\|_F^2 \\
&\overset{(c)}{\leq} 2\left(\frac{\|\mathbf{R}\|_F^2 + \|\mathbf{S}\|_F^2}{\delta_1^2} - m\right) \\
&\leq 2\left(\frac{\|\mathbf{R}\|_F^2 + \|\mathbf{S}\|_F^2}{\delta_1^2}\right), \\
&\overset{(d)}{=} \frac{4\|\mathbf{E}\|_F^2}{\delta_1^2}
\end{aligned} \tag{42}$$

where the inequality (a) is due to $\|\cdot\|_2^2 \leq \|\cdot\|_F^2$ and (b) and (c) are from Theorems 3 and 4 and respectively. (d) is due to

$$\|\mathbf{R}\|_F^2 = \|\mathbf{Q}\mathbf{S}\mathbf{V}_M - \mathbf{U}_M\mathbf{\Sigma}_M\|_F^2 = \|\mathbf{E}\mathbf{V}_M\|_F^2 = \|\mathbf{E}\|_F^2, \tag{43}$$

with a similar bound for $\mathbf{S}$. With these bounds, we prove Lemma 5 as follows

TABLE IV
SUMMARY OF THE MINIMUM VALUES OF δ_2 VARYING r AND l

$r=5,d=5,k=25$	$\lvert\sigma_r(\mathbf{QS})-\sigma_{r+1}(\hat{\mathbf{QS}})\rvert$	**min of** δ_2	$k=25,l=5,d=25$	$\lvert\sigma_r(\mathbf{QS})-\sigma_{r+1}(\hat{\mathbf{QS}})\rvert$	**min of** δ_2
$l=5$	0.1133	0.000648	$r=5$	0.1233	0.00011
$l=7$	0.4646	0.211174	$r=7$	0.116	0.00138
$l=9$	0.69002	0.460254	$r=9$	0.1083	0.00017
$l=11$	0.8054	0.562388	$r=11$	0.0971	0.00119

$$
\begin{aligned}
&\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}-\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\
&\overset{(a)}{\leq}\|\mathbf{P}_{\mathbf{U}_A}\|_2^2\left\|\mathbf{M}-\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\
&=\left\|\mathbf{M}-\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\
&=\left\|\mathbf{M}-\mathbf{M}\mathbf{P}_{\mathbf{V}_M}+\mathbf{M}\mathbf{P}_{\mathbf{V}_M}-\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\
&\overset{(b)}{\leq}2\|\mathbf{M}-\mathbf{M}\mathbf{P}_{\mathbf{V}_M}\|_2^2+2\left\|\mathbf{M}\mathbf{P}_{\mathbf{V}_M}-\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\
&\overset{(c)}{\leq}2\sigma_{r+1}^2(\mathbf{M})+2\left\|\mathbf{M}\mathbf{P}_{\mathbf{V}_M}-\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\
&\overset{(d)}{\leq}2\sigma_{r+1}^2(\mathbf{M})+2\|\mathbf{M}\|_2^2\left\|\mathbf{P}_{\mathbf{V}_M}-\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\
&\overset{(e)}{\leq}2\sigma_{r+1}^2(\mathbf{M})+2\sigma_1^2(\mathbf{M})\left\|\mathbf{P}_{\mathbf{V}_M}-\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2, \qquad (44)
\end{aligned}
$$

where the inequality (a) and (d) is due to matrix norm inequality and (b) is due to the fact that for $a,b\geq 0,(a+b)^2\leq 2(a^2+b^2)$. Inequalities (c) and (e) are derived from the operator norm property. Next, we bound $\left\|\mathbf{P}_{\mathbf{V}_M}-\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2$.

$$
\begin{aligned}
&\left\|\mathbf{P}_{\mathbf{V}_M}-\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\
&=\left\|\mathbf{P}_{\mathbf{V}_M}-\mathbf{P}_{\mathbf{V}_{QS}}+\mathbf{P}_{\mathbf{V}_{QS}}-\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\
&\overset{(a)}{\leq}2\left\|\mathbf{P}_{\mathbf{V}_M}-\mathbf{P}_{\mathbf{V}_{QS}}\right\|_2^2+2\left\|\mathbf{P}_{\mathbf{V}_{QS}}-\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2, \qquad (45)
\end{aligned}
$$

where (a) is due to the fact that for $a,b\geq 0,(a+b)^2\leq 2(a^2+b^2)$.

Finally, combining everything we have

$$
\begin{aligned}
&\left\|\mathbf{P}_{\mathbf{U}_A}\mathbf{M}-\mathbf{P}_{\mathbf{U}_A}\mathbf{M}\mathbf{P}_{\hat{\mathbf{V}}_{QS}}\right\|_2^2\\
&\leq 2\sigma_{r+1}^2(\mathbf{M})+16\sigma_1^2(\mathbf{M})\,\|\mathbf{E}\|_F^2\left(\frac{1}{\delta_1^2}+\frac{\left\|(\mathbf{S}\boldsymbol{\Psi})^\dagger\mathbf{S}\right\|_F^2}{\delta_2^2}\right) \qquad (46)
\end{aligned}
$$

⊠

APPENDIX C
NUMERICAL JUSTIFICATION OF ASSUMPTION 2

In order to address questions from anonymous reviewers from our prior submission, herein we perform extensive monte-carlo simulations and provide numerical results for the values obtained for δ_2. We provide the minimum values of δ_2 for the case of $r=5$ in Table V.

We also provide the mean and minimum values of δ_2 in Table IV for various values of l and r with a fixed k and d. Both $\mathbf{Q}$ and $\mathbf{E}$ are randomly generated, and we average over 300 realizations of $|\sigma_r(\mathbf{QS})-\sigma_{r+1}(\hat{\mathbf{QS}})|$ varying both $\mathbf{Q}$ and $\mathbf{E}$ for each case to obtain the mean. The minimum value is obtained from out of all 300 realizations for each case. Once again we see that a zero value is never attained.

TABLE V
SUMMARY OF THE MINIMUM VALUES OF $|\sigma_5(\mathbf{QS})-\sigma_6(\hat{\mathbf{QS}})|$ VARYING k WHEN $r=5$

k	Minimum value of $\lvert\sigma_5(\mathbf{QS})-\sigma_6(\hat{\mathbf{QS}})\rvert$
$k=45$	0.053367
$k=40$	0.010940
$k=35$	0.02726
$k=30$	0.01587

Finally, we note that without additional restrictive assumptions on the models, it is highly non-trivial to provide theoretical bounds on the values of δ_2. One potential method of evaluating this is to enumerate a few possible settings wherein it is easy to compute δ_2 and show that the occurance of such examples is unlikely, however a drawback of such an approach is that enumerating such pathological cases is combinatorial in nature and this prevents us from providing a rigorous theoretical analysis.

REFERENCES

[1] 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)*, 2023, pp. 1–5.

[2] 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

[3] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[4] X. Liu, X. Wang, L. Zou, J. Xia, and W. Pang, "Spatial imputation for air pollutants data sets via low rank matrix completion algorithm," *Environment international*, vol. 139, p. 105713, 2020.

[5] S. Li, L. Cheng, T. Zhang, H. Zhao, and J. Li, "Graph-guided bayesian matrix completion for ocean sound speed field reconstruction," *The Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 689–710, 2023.

[6] K. D. Harris, S. Mihalas, and E. Shea-Brown, "High resolution neural connectivity from incomplete tracing data using nonnegative spline regression," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[7] 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.

[8] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene–disease associations," *Bioinformatics*, vol. 30, no. 12, pp. i60–i68, 2014.

[9] 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.

[10] K.-Y. Chiang, I. S. Dhillon, and C.-J. Hsieh, "Using side information to reliably learn low-rank matrices from missing and corrupted observations," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 3005–3039, 2018.

[11] T. Cai, T. T. Cai, and A. Zhang, "Structured matrix completion with applications to genomic data integration," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 621–633, 2016.

[12] V. Farias, A. Li, and T. Peng, "Learning treatment effects in panels with general intervention patterns," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 001–14 013, 2021.

[13] G. Liu, Q. Liu, and X. Yuan, "A new theory for matrix completion," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] S. J. Quiton, U. Mitra, and S. Mallikarjun Sharada, "A matrix completion algorithm to recover modes orthogonal to the minimum energy path in chemical reactions," *The Journal of Chemical Physics*, vol. 153, no. 5, p. 054122, Aug. 2020. [Online]. Available: http://aip.scitation.org/doi/10.1063/5.0018326

[15] S. Bac, S. J. Quiton, K. Kron, J. Chae, U. Mitra, and S. M. Sharada, "A matrix completion algorithm for efficient calculation of quantum and variational effects in chemical reactions," *The Journal of Chemical Physics*, vol. 156, no. 18, p. 184119, 2022.

[16] 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.

[17] S. Foucart, D. Needell, R. Pathak, Y. Plan, and M. Wootters, "Weighted matrix completion from non-random, non-uniform sampling patterns," *IEEE Transactions on Information Theory*, vol. 67, no. 2, pp. 1264–1290, 2020.

[18] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1665–1697, 2012.

[19] 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.

[20] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error cur matrix decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 844–881, 2008.

[21] 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.

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

[23] M. Hašan, F. Pellacini, and K. Bala, "Matrix row-column sampling for the many-light problem," in *ACM SIGGRAPH 2007 papers*, 2007, pp. 26–es.

[24] J. Ou and F. Pellacini, "Lightslice: matrix slice sampling for the many-lights problem." *ACM Trans. Graph.*, vol. 30, no. 6, p. 179, 2011.

[25] 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.

[26] E. J. Candes and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[27] J. P. Haldar and Z.-P. Liang, "Spatiotemporal imaging with partially separable functions: A matrix recovery approach," in *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2010, pp. 716–719.

[28] Z.-P. Liang, "Spatiotemporal imagingwith partially separable functions," in *2007 4th IEEE international symposium on biomedical imaging: from nano to macro*. IEEE, 2007, pp. 988–991.

[29] B. C. Garrett and D. G. Truhlar, "Variational transition state theory. primary kinetic isotope effects for atom transfer reactions," *Journal of the American Chemical Society*, vol. 102, no. 8, pp. 2559–2570, 1980.

[30] A. Gonzalez-Lafont, T. N. Truong, and D. G. Truhlar, "Interpolated variational transition-state theory: Practical methods for estimating variational transition-state properties and tunneling contributions to chemical reaction rates from electronic structure calculations," *The Journal of Chemical Physics*, vol. 95, no. 12, pp. 8875–8894, 1991.

[31] G. Ongie, "Algebraic Variety Models for High-Rank Matrix Completion MATLAB code," Jul. 2017, (accessed 2021-04-30). [Online]. Available: https://github.com/gregongie/vmc

[32] W. H. Miller, N. C. Handy, and J. E. Adams, "Reaction path hamiltonian for polyatomic molecules," *The Journal of chemical physics*, vol. 72, no. 1, pp. 99–112, 1980.

[33] M. Page and J. W. McIver Jr, "On evaluating the reaction path hamiltonian," *The Journal of chemical physics*, vol. 88, no. 2, pp. 922–935, 1988.

[34] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

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

[36] G. W. Stewart, "Perturbation theory for the singular value decomposition," Tech. Rep., 1998.

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

[38] 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.

[39] 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.

[40] 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.

[41] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Statistical science*, vol. 27, no. 4, pp. 538–557, 2012.

[42] R. A. Horn and C. R. Johnson, "Topics in matrix analysis," 1991.

[43] G. W. Stewart, "Matrix perturbation theory," 1990.

[44] C. Rencher, Alvin and William.F, "Methods of multivariate analysis," 2012.