Journal of Machine Learning Research 24 (2023) 1-40 Submitted 3/22; Revised 4/23; Published 9/23

Near-Optimal Weighted Matrix Completion

Oscar López LOPEZO@FAU.EDU
Harbor Branch Oceanographic Institute
Florida Atlantic University
Fort Pierce, FL 34946, USA

Editor: Moritz Hardt

Abstract

Recent work in the matrix completion literature has shown that prior knowledge of a matrix's row and column spaces can be successfully incorporated into reconstruction programs to substantially benefit matrix recovery. This paper proposes a novel methodology that exploits more general forms of known matrix structure in terms of subspaces. The work derives reconstruction error bounds that are informative in practice, providing insight to previous approaches in the literature while introducing novel programs with reduced sample complexities. The main result shows that a family of weighted nuclear norm minimization programs incorporating a M_1r-dimensional subspace of $n \times n$ matrices (where $M_1 \geq 1$ conveys structural properties of the subspace) allow accurate approximation of a rank r matrix aligned with the subspace from a near-optimal number of observed entries (within a logarithmic factor of M_1r). The result is robust, where the error is proportional to measurement noise, applies to full rank matrices, and reflects degraded output when erroneous prior information is imposed. Numerical experiments are presented that validate the theoretical behavior derived for several example weighted programs.

Keywords: Matrix completion, weighted matrix completion, low rank matrices, incoherence, nuclear norm minimization, convex optimization

1. Introduction

Over the past two decades, matrix completion has evolved from an academic curiosity to a common industrial tool (Recht et al. 2010; Candès and Recht 2009; Srebro and Jaakkola 2004; Foygel and Srebro 2011). Its utility includes seismic data acquisition (Aravkin et al. 2014; López et al. 2016; Yang et al. 2013), machine learning (Nguyen et al. 2018; Yi et al. 2012; Luo et al. 2015), collaborative filtering (Srebro and Salakhutdinov 2010), computer vision (Tomasi and Kanade 1992), gene expression analysis (Troyanskaya et al. 2001) and MRI (Zhao et al. 2010). In these applications, practitioners wish to estimate a data matrix of interest $D \in \mathbb{C}^{n_1 \times n_2}$ from a fraction of revealed noisy entries. The success of recent approaches hinges on the underlying assumption that D can be well approximated by a rank r matrix where $r \ll \min\{n_1, n_2\}$, that is, D has *low rank structure.* This data model is common in smooth signals, but pervasive simply by the sheer nature of large scale data (Udell and Townsend 2019).

To elaborate, let $\Omega \subseteq \{1, \cdots, n_1\} \times \{1, \cdots, n_2\}$ be a subset of size $m \leq n_1 n_2$ and $P_\Omega : \mathbb{C}^{n_1 \times n_2} \mapsto \mathbb{C}^m$ the corresponding *sampling operator* that extracts the m values at the entries specified by Ω from an input matrix (with $|\Omega| = m$). Given $P_\Omega(D) + d \in \mathbb{C}^m$, where

$d \in \mathbb{C}^m$ encompasses measurement noise, the goal of matrix completion is to recover D as accurately as possible. Under the low rank assumption, a well-studied method to estimate the data matrix is via the *nuclear norm minimization* program (Fazel et al. 2001; Recht et al. 2010; Candès and Recht 2009), which estimates D via

$$D^1 := \underset{X \in \mathbb{C}^{n_1 \times n_2}}{\operatorname{argmin}} \ \|X\|_* \ \text{ subject to } \ \|P_\Omega(D) + d - P_\Omega(X)\|_2 \leq \eta, \tag{1}$$

where $\|X\|_* = \sum_{k=1}^{\min(n_1,n_2)} \sigma_k(X)$ is the nuclear norm, $\sigma_k(X)$ is the k-th largest singular value of X and η is a program parameter chosen according to the noise level. The nuclear norm penalty provides a convex surrogate for the rank objective function, in order to output a low rank matrix that is viable for the noisy observations in a tractable manner.

As in previous work, a notion of *incoherence* is needed to quantify how evenly distributed the information is throughout the data matrix. Ultimately, incoherence conditions ensure that a set of observed entries Ω chosen uniformly at random provide an appropriate sampling scheme for the array of interest.

Definition 1 *Given $D \in \mathbb{C}^{n_1 \times n_2}$ and $r \leq \min\{n_1, n_2\}$, consider the singular value decomposition (SVD) $D = U\Sigma V^\top$. The r-*incoherence parameters *of D are defined as the smallest $\mu_0, \mu_1 > 0$ such that*

$$\max_{1\leq k\leq n_1} \sqrt{\sum_{j=1}^{r} \left|U_{kj}\right|^2} \leq \sqrt{\frac{\mu_0 r}{n_1}}, \qquad \max_{1\leq \ell\leq n_2} \sqrt{\sum_{j=1}^{r} \left|V_{\ell j}\right|^2} \leq \sqrt{\frac{\mu_0 r}{n_2}} \tag{2}$$

and

$$\max_{k,\ell} \sqrt{\sum_{j=1}^{r} \left|U_{kj}\right|^2 \left|V_{\ell j}\right|^2} \leq \sqrt{\frac{\mu_1 r}{n_1 n_2}}. \tag{3}$$

Definition (2) is common in the literature, know as the *standard incoherence* condition (Chen 2015). The parameter μ_1 is unique to this work, but similar to the joint incoherence condition introduced in Recht 2011. This novel parameter will be elaborated in Section 3.2. Intuitively, small parameters (for example, $\mu_0, \mu_1 \sim \log(\max\{n_1, n_2\})$) correspond to data matrices whose information is not concentrated on a few set of entries. Such metrics of "spikiness" are necessary when the observations are chosen without regard to the matrix structure, in order to guarantee that the probed entries will supply a substantial amount of information. However, incoherence conditions can be avoided if the sampling scheme is modified according to prior knowledge of the matrix's leverage scores (Chen et al. 2015, 2014; Eftekhari et al. 2018a).

The following result states the sample complexity and resulting error bound for program (1), where without loss of generality it is henceforth assumed that $n_1 \geq n_2$.

Theorem 2 *Let $D \in \mathbb{C}^{n_1 \times n_2}$ have r-incoherence parameters μ_0, μ_1 and suppose $\Omega \subseteq [n_1] \times [n_2]$ is generated by selecting a subset of size $m \leq n_1 n_2$ uniformly at random from all subsets of size m. Define D^1 as in (1) with $\|d\|_2 \leq \eta$. There exist universal constants $c_0, c_1, c_2 > 0$ such that if*

$$m \geq c_0 \max\{\mu_0, \sqrt{\mu_1}\} n_1 r \log^2(n_1)$$

then with high probability

$$\|D - D^1\|_F \leq c_1 \sqrt{\frac{n_1 n_2 \log(n_1)}{m}} \sum_{k=r+1}^{n_2} \sigma_k(D) + c_2 \frac{n_1 n_2 \sqrt{r} \log(n_1)}{m} \eta.$$

The result matches previous work in terms of sample complexity required for exact matrix completion (Chen 2015; Eftekhari et al. 2018a,b). However, a fair comparison is difficult to make due to the dependence here on $\sqrt{\mu_1}$ whereas other authors only require μ_0. Section 3.2 will provide data matrices with $\sqrt{\mu_1} < \mu_0$ as well as examples where $\sqrt{\mu_1} \geq \mu_0$ holds. Therefore, there is no strict relationship between these parameters and Theorem 2 is arguably on par with incoherence-optimal conditions (Chen 2015). These results state that, when a practitioner is oblivious to the data matrix's structure, $n_1 r \log^2(n_1)$ observed entries are needed to robustly reconstruct an incoherent $n_1 \times n_2$ rank r matrix.

1.1 Matrix Completion with Prior Knowledge: Approach and Overview of the Main Results

In many applications, prior knowledge of the data's structure is available. Incorporating this information appropriately into a reconstruction program has been shown to significantly improve the success of matrix recovery (Aravkin et al. 2014; Zhang et al. 2019, 2020; Abernethy et al. 2009; Bayat and Daei 2020; Chiang et al. 2015; Eftekhari et al. 2018b; Chen 2015; Xu et al. 2013; Yi et al. 2013; Jain and Dhillon 2013; Chen et al. 2015, 2014). Inspired by these approaches, this paper proposes and analyzes a matrix reconstruction framework that exploits prior knowledge of subspaces that align well with the matrix of interest. This section introduces the approach and summarizes the results and novelties of this paper.

The contribution of this work is in the generality of the proposed framework and analysis. The main result provides the ability to derive near-optimal sample complexities and informative error bounds that express a trade-off when incorporating distinct subspaces enforcing known matrix structure. The result applies to a variety of *weighted nuclear norm minimization* programs, providing novel insight to previous approaches in the literature and proposing new programs.

To introduce the approach and summarize the results, let $T \subset \mathbb{C}^{n_1 \times n_2}$ be a linear subspace of matrices with orthogonal complement $T^\perp$ and respective orthogonal projections $\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$. With weight parameter $0 \leq \omega \leq 1$, the family of weighted nuclear norm minimization programs proposed here approximate the data matrix via the following modified version of (1)

$$D^\omega \coloneqq \operatorname*{argmin}_{X \in \mathbb{C}^{n_1 \times n_2}} \left\| \omega \mathcal{P}_T(X) + \mathcal{P}_{T^\perp}(X) \right\|_* \text{ subject to } \|P_\Omega(D) + d - P_\Omega(X)\|_2 \leq \eta. \quad (4)$$

When $\omega < 1$, program (4) favors matrices that align with the estimate subspace T while the case $\omega = 1$ reduces the program to unbiased nuclear norm minimization (1). The weight parameter toggles how severely one wishes to penalize matrices that do not agree with the prior information, thereby capturing the user's confidence in T. The main novelty of program (4) in contrast to previous approaches is in the ability to incorporate subspaces with general structure and the flexibility of weight selection.

The theoretical contributions of this paper can be summarized as follows:

- Theorem 4 analyzes program (4) in a general sense, applying to any subspace T with elements of maximal rank r. The main result states that one can accurately reconstruct a matrix nearly lying in T from $m \geq rM_1 \operatorname{polylog}(n_1)$ observed entries, where $1 \leq M_1 \leq \frac{n_1 n_2}{r}$ captures crucial dimensional and incoherence-based properties of T. The result is robust, with error bound proportional to the measurement noise level η, the error of the best rank r approximation $\sum_{k=r+1}^{n_2} \sigma_k$, and a term that quantifies the accuracy of T.

- Specific choices of T are presented in Section 2.1, demonstrating the applicability of Theorem 4. The results derived therein showcase a variety of sample complexities including $m \sim r \operatorname{polylog}(n_1)$, $m \sim r^2 \operatorname{polylog}(n_1)$, and $m \sim (n_1 r - r^2) \operatorname{polylog}(n_1)$. Furthermore, the derived error bounds express an informative trade-off between sample complexity and sensitivity to the accuracy of T. In other words, programs incorporating subspaces T that require less samples will in general exhibit error terms that are more susceptible to inaccurate T (quantified via the principal angles between subspaces, Ji-guang 1987). This behavior is validated numerically in Section 4.

- Some examples in Section 2.1 are related to approaches previously studied in the literature: Yi et al. 2013; Bayat and Daei 2020; Chiang et al. 2015; Eftekhari et al. 2018b; Chen 2015; Xu et al. 2013; Jain and Dhillon 2013. Most of the methodologies or results from these citations only apply in a high fidelity scenario, when T is error-free or aligns sufficiently well with the data matrix. In contrast, the main contribution here is the robustness of program (4) and the theoretical results. The error bounds derived here provide novel insight to previous approaches, allowing inexact prior information while roughly matching the sample complexity of related results in the literature. See Section 3 for further discussion.

1.2 Organization and Notation

The remainder of the paper is organized as follows: Section 2 discusses a foundational weighted matrix completion approach from the literature in order to elaborate on the inspiration for this work, its main result, and the improvements provided relative to the literature. Section 2.1 applies the main result to example subspaces T, deriving a variety of sample complexities and error bounds. Section 3 discusses related work in the literature and the introduced incoherence parameter μ_1 in order to fairly compare this work with other results. Section 4 conducts numerical experiments, comparing example programs to the original weighted program discussed in Section 2. The paper concludes with a discussion of future work in Section 5 followed by the proofs in the Appendix.

Notation: for any integer $n \in \mathbb{N}$, $[n]$ denotes the set $\{\ell \in \mathbb{N} : 1 \leq \ell \leq n\}$ and I_n is the $n \times n$ identity matrix. For $k, \ell \in \mathbb{N}$, b_k indicates the k-th entry of the vector b, $X_{k\ell}$ denotes the (k, ℓ) entry of the matrix X and X_{k*} ($X_{*\ell}$) denotes its k-th row (resp. ℓ-th column). For vectors, $\|b\|_2$ is the Euclidean norm. For matrices, $\sigma_k(X)$ denotes the k-th largest singular value of X, $\|X\| := \sigma_1(X)$ is the operator norm, $\|X\|_F := \langle X, X \rangle^{1/2}$ is the Frobenius norm, $\|X\|_* := \sum_k \sigma_k(X)$ is the nuclear norm, and $\|X\|_\infty$ is the largest entry of X in absolute value. S and S_{op} are the closed unit balls in $\mathbb{C}^{n_1 \times n_2}$ with respect to the Frobenius and operator norms respectively. The adjoint of a linear operator $\mathcal{A}$ is denoted by $\mathcal{A}^*$, while $X^\top$

will be used to denote the conjugate transpose of a matrix X. As previously mentioned, for matrices $X \in \mathbb{C}^{n_1 \times n_2}$ it is assumed that $n_1 \geq n_2$ without loss of generality.

2. Weighted Matrix Completion

To the author's best knowledge, the first version of a weighted nuclear norm minimization program was proposed by Aravkin et al. 2014. To elaborate on this original approach, given $r \leq n_2$, consider the SVD and decompose the data matrix of interest as

$$D = U\Sigma V^\top = U^r \Sigma^r V^{r\top} + U^+ \Sigma^+ V^{+\top}$$

where $U^r \Sigma^r V^{r\top}$ pertains to the largest r singular values of D along with the corresponding singular vectors. Assume that $\tilde{U} \in \mathbb{C}^{n_1 \times r}, \tilde{V} \in \mathbb{C}^{n_2 \times r}$ with orthonormal columns are available containing information of the range of $U^r \in \mathbb{C}^{n_1 \times r}, V^r \in \mathbb{C}^{n_2 \times r}$ and define

$$Q_{\omega_1} := \omega_1 \tilde{U}\tilde{U}^\top + I_{n_1} - \tilde{U}\tilde{U}^\top, \qquad W_{\omega_2} := \omega_2 \tilde{V}\tilde{V}^\top + I_{n_2} - \tilde{V}\tilde{V}^\top,$$

where $\omega_1, \omega_2 \in [0,1]$ are chosen weights. Notice that Q_1, W_1 are identity matrices and otherwise, when $\omega_1, \omega_2 < 1$, these linear operators skew toward the orthogonal complement of range$(\tilde{U})$ and range$(\tilde{V})$ respectively. The original weighted nuclear norm minimization program approximates the data matrix via

$$D^{\omega_1,\omega_2} := \underset{X \in \mathbb{C}^{n_1 \times n_2}}{\operatorname{argmin}} \|Q_{\omega_1} X W_{\omega_2}\|_* \ \text{ subject to } \ \|P_\Omega(D) + d - P_\Omega(X)\|_2 \leq \eta. \tag{5}$$

Analogous to this paper's approach, with $\omega_1, \omega_2 < 1$ program (5) favors matrices that match a certain structure while $\omega_1 = \omega_2 = 1$ is unbiased nuclear norm minimization (1). Notice that (5) is nearly of the form (4), but the choice of two weights in the original formulation will impede the results here from being directly applicable. However, a program of the form (4) that is closely related to (5) will be considered in Section 2.1.

Spurring from the original program, similar methodologies have been proposed in the literature (Eftekhari et al. 2018b; Zhang et al. 2019, 2020; Bayat and Daei 2020) but distinct approaches have also been considered (Xu et al. 2013; Chiang et al. 2015; Yi et al. 2013; Jain and Dhillon 2013; Abernethy et al. 2009; Chen 2015). To attempt producing a result that provides some level of insight for many of these variations, a more general notion of incoherence that applies to an entire subspace is required.

Definition 3 *For a subspace $T \subset \mathbb{C}^{n_1 \times n_2}$ and $\rho \leq n_2$, the subspace ρ-incoherence and joint incoherence parameters of T are defined respectively as*

$$M_0 := \max_{X \in T \cap S_{op}} \frac{n_2}{\rho} \|X\|_{\infty,2}^2 \tag{6}$$

and

$$M_1 := \max_{X \in T \cap S} \frac{n_1 n_2}{\rho} \max_{k,\ell} |X_{k\ell}|^2, \tag{7}$$

where $\|X\|_{\infty,2}$ is the maximum of the row and column norms of X, see (24).

The ensemble will be referred to as the *ρ-subspace incoherence* parameters or condition. These parameters will provide crucial dimensional information of a given subspace T. Section 2.1 will explore these parameters via example programs, with results that demonstrate their role in deriving near-optimal sample complexities.

Henceforth, let ρ be defined as

$$\rho = \max_{X \in T} \ \mathrm{rank}(X). \tag{8}$$

The main result of the paper can now be presented:

Theorem 4 *Let $T \subset \mathbb{C}^{n_1 \times n_2}$ be a subspace with ρ-subspace incoherence parameters M_0, M_1. Suppose $\Omega \subseteq [n_1] \times [n_2]$ is generated by selecting a subset of size $m \leq n_1 n_2$ uniformly at random from all subsets of size m. Let $D \in \mathbb{C}^{n_1 \times n_2}$ and define D^ω as in (4) with $\|d\|_2 \leq \eta$. There exist universal constants $c_0, c_1, c_2, c_3, c_4 > 0$ such that if*

$$m \geq c_0 M_1 \rho \log^3(n_1) \qquad \textit{and} \qquad 0 < \omega \leq \frac{\sqrt{M_1}\log(n_1+n_2)}{\sqrt{n_1 n_2}} \tag{9}$$

or

$$m \geq c_0 \max\left\{M_0, \sqrt{M_1}\right\}\omega n_1 \rho \log^2(n_1) \qquad \textit{and} \qquad \frac{1}{2\sqrt{2\rho\log(n_1)}} \leq \omega \leq 1, \tag{10}$$

then

$$\begin{aligned} \|D - D^\omega\|_F &\leq c_1 f(\omega)\left(\sqrt{\frac{n_1 n_2 \log(n_1)}{m}} + \frac{1}{\omega}\right)\sum_{k=\rho+1}^{n_2}\sigma_k(D) \\ &+ c_2 f(\omega)\left(\frac{1}{\omega} + \sqrt{\frac{n_1 n_2 \log(n_1)}{m}}\left(\omega\sqrt{\frac{n_1 n_2 \rho \log(n_1)}{m}} + 1\right)\right)\eta \\ &+ c_3 f(\omega)\left(\frac{1}{\omega} + \sqrt{\frac{n_1 n_2 \log(n_1)}{m}}\right)\left\|\mathcal{P}_{T^\perp}(U^\rho \Sigma^\rho V^{\rho\top})\right\|_* \end{aligned} \tag{11}$$

holds with probability greater than $1 - \frac{c_4}{n_1}$, where

$$f(\omega) := \min\left\{\omega\sqrt{\frac{n_1 n_2 \log(n_1)}{m}}, 1\right\}.$$

As in the beginning of the section, $U^\rho \Sigma^\rho V^{\rho\top}$ in (11) denotes the best rank ρ approximation of D via the SVD. The proof is postponed until Section A.3. The theorem implies that faithful prior subspace knowledge (with elements of maximal rank ρ) can be used to robustly approximate a nearly rank ρ matrix in the subspace from as few as $M_1 \rho \log^3(n_1)$ observed entries. The result applies to two intervals for ω, with small and large weights ($\omega \approx 0$ and $\omega \approx 1$ respectively). These ranges offer distinct sample complexities and error bounds. The main focus of this paper will be in the regime of small weights, which produce a substantial reduction in sample complexity. However, the interval of larger weights (10) also generates interesting results (see end of Section 2.1).

The result is robust to inexact subspaces, expressed in the term

$$\left\|\mathcal{P}_{T^\perp}(U^\rho \Sigma^\rho V^{\rho\top})\right\|_*$$

appearing in (11). This term quantifies the inaccuracy of the estimate subspace relative to the best rank ρ approximation of the data, where $\|\mathcal{P}_{T^\perp}(U^\rho \Sigma^\rho V^{\rho\top})\|_* \approx 0$ implies that T was chosen appropriately. Arguably, this term is quite raw and the result can be more informative if a more general metric is used. The main convenience in presenting Theorem 4 with error bound (11) is that no restrictions are imposed on T.

To provide a more informative error bound (at the cost of restricting T), the accuracy of the estimate subspace can instead be quantified via the *principal angle between subspaces* (PABS) Drmac 2000

$$\sin(\theta_T) := \left\|\mathcal{P}_{T^\perp}\mathcal{P}_{T_\rho}\right\|_{F\to F}. \tag{12}$$

In (12), $\|\circ\|_{F\to F}$ is the spectral norm for operators acting on matrices and $\mathcal{P}_{T_\rho}$ is the orthogonal projection onto a ρ-dimensional subspace of matrices that $U^\rho \Sigma^\rho V^{\rho\top}$ belongs to

$$T_\rho := \operatorname{span}\left\{U^\rho_{*k} V^{\rho\top}_{*k}\right\}_{k=1}^{\rho}.$$

Notice that the PABS lies in $[0,1]$, where $\sin(\theta_T) \approx 0$ implies accurate subspace estimate and $\sin(\theta_T) \approx 1$ states that $U^\rho \Sigma^\rho V^{\rho\top}$ largely lies in $T^\perp$. Since $\sin(\theta_T) \neq 1$ only when $\dim(T) \geq \dim(T_\rho)$ (see Drmac 2000), this restriction on T must now be enforced in order to obtain the following modified version of Theorem 4 involving the PABS:

Corollary 5 *Under the same conditions of Theorem 4, if* $\dim(T) \geq \rho$ *then*

$$\begin{aligned}
\|D - D^\omega\|_F \leq{}& c_1 f(\omega)\left(\sqrt{\frac{n_1 n_2 \log(n_1)}{m}} + \frac{1}{\omega}\right)\sum_{k=\rho+1}^{n_2}\sigma_k(D) \\
&+ c_2 f(\omega)\left(\frac{1}{\omega} + \sqrt{\frac{n_1 n_2 \log(n_1)}{m}}\left(\omega\sqrt{\frac{n_1 n_2 \rho \log(n_1)}{m}} + 1\right)\right)\eta \\
&+ c_3 f(\omega)\sin(\theta_T)\left(\frac{1}{\omega} + \sqrt{\frac{n_1 n_2 \log(n_1)}{m}}\right)\left(n_2 \sum_{k=1}^{\rho}\sigma_k^2(D)\right)^{1/2}
\end{aligned}$$

holds with probability greater than $1 - \frac{c_4}{n_1}$.

Corollary 5 provides a more standard error bound, since the PABS are well understood metrics of signal alignment in the literature. However, aside from imposing $\dim(T) \geq \rho$, notice that the last term in the error bound above now includes the avoidable factor $\sqrt{n_2}$. For this reason, the example results of Section 2.1 will apply Theorem 4 rather than Corollary 5 since the latter seems to produce pessimistic error bounds in general.

To further understand the implications of the main results and demonstrate the crucial role of the subspace incoherence parameters, it is instructive to consider specific choices of T. This is the purpose of Section 2.1, where sample complexities will be supplied for various estimate subspaces. The proofs of these results, which are corollaries of Theorem 4, consist of lower and upper bounding M_1. The lower bounds (presented in Appendix B) enforce the sample complexity upholding the dimensionality of T. These examples illustrate the importance of M_1 and the near optimality of the main result.

2.1 Matrix Completion with Prior Knowledge: Example Subspaces

This section provides example programs that enforce specific subspace structure on the output estimate data matrix. Theorem 4 will be applied in the small weight regime (9) to derive informative reconstruction error bounds and nearly optimal sample complexities. In particular, the results will showcase a trade-off when incorporating distinct choices of T. The examples will illustrate that subspaces allowing exact matrix completion from relatively few samples will in general exhibit increased sensitivity to inaccurate prior information.

To elaborate in a concrete setting, decompose as before $D = U\Sigma V^\top = U^r\Sigma^r V^{r\top} + U^+\Sigma^+ V^{+\top}$, where $U^r\Sigma^r V^{r\top}$ is the best rank r approximation of D. Assume that $\tilde{U} \in \mathbb{C}^{n_1\times r}, \tilde{V} \in \mathbb{C}^{n_2\times r}$ with orthonormal columns are available containing information of U^r, V^r. In this section, $\tilde{U}$ and $\tilde{V}$ will specify T. The accuracy of the prior information will be mainly quantified via the principal angle between subspaces (PABS) with respect to the range of the left and right singular vectors. From the SVD, notice that the columns of U^+ (V^+ respectively) form an orthonormal basis for range$(U^r)^\perp$ (range$(V^r)^\perp$ respectively). The PABS here are defined via the canonical correlation coefficients (Ji-guang 1987)

$$\sin(\theta_u) := \left\|\tilde{U}^\top U^+\right\| \quad \text{and} \quad \sin(\theta_v) := \left\|\tilde{V}^\top V^+\right\|, \tag{13}$$

where $\|X\| = \sigma_1(X)$ denotes the operator norm of a matrix X. The PABS lie in $[0,1]$ and provide a measure of the degree of alignment between the data's main rank r component and estimates specified in some sense by $\tilde{U}$ and $\tilde{V}$. The case $\sin(\theta_u), \sin(\theta_v) \approx 0$ holds with accurate knowledge, while $\sin(\theta_u), \sin(\theta_v) \approx 1$ implies inappropriate prior information was supplied.

With this in mind, simplified results for four programs of the form (4) with specific choices of T are provided. Henceforth, program (4) incorporating a subspace T will be refered to as program $(4, T)$. The following examples are presented in order of most to least sensitive error bounds with respect to the inaccuracy of T, respectively exhibiting a trend of increasing sample complexities.

Theorem 6 *Let $D \in \mathbb{C}^{n_1\times n_2}$ have rank r, and consider estimates $\tilde{U} \in \mathbb{C}^{n_1\times r}$ and $\tilde{V} \in \mathbb{C}^{n_2\times r}$ where $\tilde{U}\tilde{V}^\top$ has r-incoherence parameter μ_1. Suppose $\Omega \subseteq [n_1]\times[n_2]$ is generated by selecting a subset of size $m \le n_1 n_2$ uniformly at random from all subsets of size m. Define D^ω as in (4) with $\|d\|_2 = \eta = 0$ and estimate subspace*

$$T_1 := \mathrm{span}\left\{\tilde{U}_{*k}\tilde{V}_{*k}^\top\right\}_{k\in\{1,\cdots,r\}}.$$

There exists universal constants $c_0, c_3 > 0$ such that if

$$m \ge c_0\mu_1 r\log^3(n_1) \quad \text{and} \quad 0 < \omega \le \frac{\sqrt{\mu_1}\log(n_1+n_2)}{\sqrt{n_1 n_2}}$$

then with high probability

$$\frac{\|D - D^\omega\|_F}{\|D\|_F} \le c_3\sqrt{\frac{n_1 n_2 r\log(n_1)}{m}}\left(\sin(\theta_u) + \sin(\theta_v) + \left(\sum_{k\ne\ell}\left\|U^{r\top}\tilde{U}_{*k}\tilde{V}_{*\ell}^\top V^r\right\|_F^2\right)^{1/2}\right). \tag{14}$$

To the author's best knowledge, program (4, T_1) provides a novel method by which to exploit prior information of the data's rank 1 components. The result states that, with an appropriate weight and accurate projection onto the subspace spanned by data's rank 1 components, program (4, T_1) can faithfully estimate a rank r matrix with $\mathcal{O}(r\log^3(n_1))$ random samples. In a high fidelity scenario, there are roughly r degrees of freedom to approximate $U^r\Sigma^r V^{r\top}$. Therefore the derived sample complexity is within a quadratic logarithmic factor of the optimal rate $\mathcal{O}(r\log(n_1))$, where the logarithmic factor in the optimal rate is unavoidable in matrix completion under random sampling models (see for example Theorem 1.7 provided by Candes and Tao 2010). On the other hand, in the case of unreliable information, the right hand side of (14) showcases high sensitivity to inaccurate T_1. Several elaborated remarks are in order:

- Incorporating accurate subspace T_1 allows for frugal completion of matrices that do not satisfy typical incoherence conditions. For example, with severe parameter $\mu_0 \sim \sqrt{n_1}/r$, since $\mu_1 \leq \mu_0^2 r$ Theorem 6 guarantees reconstruction from $\mathcal{O}(n_1\log^3(n_1))$ random samples. Without prior information, such data matrices require a significant amount of additional samples to be recovered according to Theorem 2 and similar results in the literature.

- Program (4, T_1) allows reconstruction of general full rank matrices in an underdetermined scenario with prior information. With n_2 trustworthy rank 1 components (or an orthogonal projection onto the span), the result allows for an accurate estimate of a full rank matrix from $\mathcal{O}(n_2\log^3(n_1))$ observed entries.

- In contrast to other approaches that will be considered, program (4, T_1) is most sensitive to inaccurate prior information. This observation will become clear due to the final term in (14) involving the sum over all $k \neq \ell$, which is not present in the remaining error bounds. This term requires that each matrix $\tilde{U}_{*k}\tilde{V}_{*\ell}^\top$ with $k \neq \ell$ be unaligned with elements in

$$\operatorname{span}\left\{U^r_{*k}V^{r\top}_{*\ell}\right\}_{k,\ell\in\{1,\cdots,r\}}.$$

  This requisite is quite strict, since even a single inaccurate rank-1 component included as prior knowledge can cause a significant error according to (14).

The next result involves a methodology related to many previously proposed in the literature (Chen 2015; Xu et al. 2013; Chiang et al. 2015; Yi et al. 2013; Jain and Dhillon 2013; Abernethy et al. 2009). The result will render a less sensitive methodology at the cost of higher sample complexity.

Theorem 7 *Under the same setup as Theorem 6, let $\tilde{U}\tilde{U}^\top$ and $\tilde{V}\tilde{V}^\top$ have r-standard incoherence parameters μ_L and μ_R respectively. Define D^ω as in (4) with $\eta = 0$ and subspace estimate*

$$T_2 := \operatorname{span}\left\{\tilde{U}_{*k}\tilde{V}_{*\ell}^\top\right\}_{(k,\ell)\in\{1,\cdots,r\}\times\{1,\cdots,r\}}.$$

There exists universal constants $c_0, c_3 > 0$ such that if

$$m \geq c_0\mu_L\mu_R r^2\log^3(n_1) \quad \textit{and} \quad 0 < \omega \leq \frac{\sqrt{\mu_L\mu_R r}\log(n_1+n_2)}{\sqrt{n_1 n_2}}$$

then with high probability

$$\frac{\|D - D^{\omega}\|_F}{\|D\|_F} \leq c_3 \left(\sin(\theta_u) + \sin(\theta_v)\right) \sqrt{\frac{n_1 n_2 r \log(n_1)}{m}}. \tag{15}$$

This result considers a robust approach in contrast to program (4, T_1), with a more lenient requisite that the columns of U^r and V^r lie in the range of $\tilde{U}$ and $\tilde{V}$ respectively. This comparison is evident in the produced error bound (15), which is similar to (14) but no longer exhibits the sensitive term that sums over $k \neq \ell$. However, the trade-off is an increased sample complexity of $\mathcal{O}(r^2 \log^3(n_1))$. When $\omega \approx 0$, this methodology and result roughly agree with the work by Yi et al. 2013; Chen 2015; Xu et al. 2013; Jain and Dhillon 2013, but program (4, T_2) is more flexible since it introduces a choice of weight and allows for erroneous row and column subspaces (for further discussion on related work see Section 3.1).

Theorems 6 and 7 help illustrate the crucial role of the ρ-subspace incoherence parameters from Definition 3. Consider program (4, T_2), where $\rho = r$. To obtain Theorem 7 from Theorem 4, it will be shown in Appendix B that $M_1(T_2) = \mu_L \mu_R r$. Since T_2 is an r^2 dimensional space, one should not expect exact matrix recovery via program (4, T_2) with less than $r^2 \log(n_1)$ randomly sampled entries. This sample complexity is enforced by the lower bound $\mu_L \mu_R r \geq r$, which illustrates the near optimality of the main result. Similarly, for program (4, T_1) it holds that $M_1(T_1) = \mu_1 \geq 1$ which upholds the dimensionality of T_1.

The next example involves only incorporating row or column span information.

Theorem 8 *Under the same setup as Theorem 6, let $\tilde{U}\tilde{U}^{\top}$ have r-standard incoherence parameter μ_L. Define D^{ω} as in (4) with $\eta = 0$ and subspace estimate*

$$T_3 := \left\{ X \in \mathbb{C}^{n_1 \times n_2} \mid X = \tilde{U}\tilde{U}^{\top} X \right\}.$$

There exists universal constants $c_0, c_3 > 0$ such that if

$$m \geq c_0 \mu_L n_2 r \log^3(n_1) \quad \text{and} \quad 0 < \omega \leq \frac{\sqrt{\mu_L} \log(n_1 + n_2)}{n_1}$$

then with high probability

$$\frac{\|D - D^{\omega}\|_F}{\|D\|_F} \leq c_3 \sin(\theta_u) \sqrt{\frac{n_1 n_2 r \log(n_1)}{m}}. \tag{16}$$

In contrast to unbiased nuclear norm minimization (1), the result demonstrates that one sided information can reduce the sample complexity for rectangular matrices with $n_2 \ll n_1$. Analogously, for wide matrices one can incorporate information of the range instead. Comparing to the previous examples (incorporating T_1 and T_2), notice that error bound (16) is less sensitive to inaccurate subspaces as it only involves a single PABS term.

The final example attempts to provide some intuition for the original weighted nuclear norm approach. Although Theorem 4 is not directly applicable to program (5) and variations, some intuition may be provided by considering the program of the form (4) with estimate subspace

$$T_4 := \left\{ X \in \mathbb{C}^{n_1 \times n_2} \mid X = \tilde{U}\tilde{U}^{\top} X \tilde{V}\tilde{V}^{\top} + \tilde{U}\tilde{U}^{\top} X \tilde{V}^{\perp}\tilde{V}^{\perp\top} + \tilde{U}^{\perp}\tilde{U}^{\perp\top} X \tilde{V}\tilde{V}^{\top} \right\} \tag{17}$$

where the orthonormal columns of $\tilde{U}^\perp$ (resp. $\tilde{V}^\perp$) span range$(\tilde{U})^\perp$ (resp. range$(\tilde{V})^\perp$). Arguably, among all programs considered here, this approach is most related to the original weighted program. To gain insight, a new notion of incoherence is needed.

Definition 9 *Given $D \in \mathbb{C}^{n_1 \times n_2}$ and $r \leq \min\{n_1, n_2\}$, consider the singular value decomposition (SVD) $D = U\Sigma V^\top$. The r-*complementary incoherence parameter* of D is defined as the largest μ_2 such that*

$$\min_{1\leq k\leq n_1} \sqrt{\sum_{j=1}^{r} \left|U_{kj}\right|^2} \geq \sqrt{\frac{\mu_2 r}{n_1}}, \qquad \min_{1\leq \ell\leq n_2} \sqrt{\sum_{j=1}^{r} \left|V_{\ell j}\right|^2} \geq \sqrt{\frac{\mu_2 r}{n_2}}. \tag{18}$$

This incoherence condition is strongly related with the standard incoherence parameter. Notice that $1 = \|\tilde{U}_{p*}\|_2^2 + \|\tilde{U}_{p*}^\perp\|_2^2$ for any $p \in [n_1]$ and the same observation holds for $\tilde{V}$. Therefore, $0 \leq \mu_2 \leq 1 \leq \mu_0$ and for this reason the new condition is referred to as complementary. This new parameter also quantifies how spiky a data matrix is, where $\mu_2 \approx 0$ holds for matrices with entire rows or columns containing very little information and $\mu_2 \approx 1$ implies an even distribution of non-zero entries throughout the matrix. The result for the weighted program incorporating (17) now follows.

Theorem 10 *Under the same setup as Theorem 6, let $\tilde{U}\tilde{V}^\top$ have r-standard and complementary incoherence parameters μ_0 and μ_2 (resp.). Define D^ω as in (4) with $\eta = 0$ and estimate subspace T_4 from (17).*

There exists universal constants $c_0, c_3 > 0$ such that if

$$m \geq c_0\mu_0 \max\{\mu_0 r, n_1 - \mu_2 r\} r \log^3(n_1) \quad \text{and} \quad 0 < \omega \leq \frac{\sqrt{n_1 - \mu_2 r}\log(n_1+n_2)}{\sqrt{n_1 n_2}}$$

then with high probability

$$\frac{\|D - D^\omega\|_F}{\|D\|_F} \leq c_3 \sin(\theta_u)\sin(\theta_v)\sqrt{\frac{n_1 n_2 r \log(n_1)}{m}}. \tag{19}$$

Under lenient conditions, it holds that $\max\{\mu_0 r, n_1 - \mu_2 r\} = n_1 - \mu_2 r$ and therefore a $n_1 \times n_2$ rank r matrix can be recovered from $\sim \mu_0(n_1 r - \mu_2 r^2)\log^3(n_1)$ observed entries with prior information. Notice that T_4 is an $n_1 r + n_2 r - r^2$ dimensional subspace, which is respected by the complexity of the result. In contrast to unbiased nuclear norm minimization, this approach provides a slight reduction in sample complexity by requiring $\mu_0\mu_2 r^2 \log^3(n_1)$ less samples. However, the current result exhibits a larger logarithmic dependency while removing the incoherence condition related to μ_1, so a fair comparison is difficult to make.

Relative to the other approaches considered in this section, program (4, T_4) is most robust to inaccurate prior information since it imposes less of a constraint on the search space. This is exhibited in (19) via the term $\sin(\theta_u)\sin(\theta_v)$, which is the smallest error bound in the examples thusfar. However, this reduction in sensitivity to noise produces one of the largest sampling requisites of this section. The trend expressed by these error bounds will be explored numerically in Section 4, where decreasing susceptibility to inaccurate prior information in general requires an increasing number of observed entries for exact matrix completion.

The results in this section all apply Theorem 4 in the regime of small weights (9). The context of larger weights (10) can also supply interesting results. For example, program (4, T_1) with $\omega = (r\log(n_1))^{-1/2}$ can be shown to require $\max\{\mu_0, \sqrt{\mu_1}\}n_1\sqrt{r}\log^{3/2}(n_1)$ samples for accurate matrix estimation. Similarly, program (4, T_2) would need $\mu_0 n_1 r\log^{3/2}(n_1)$ observed entries, mildly reducing the sample complexity in contrast to (1) by a logarithmic factor. Intuitively, such larger weight choices would be most appropriate for less reliable prior information while still reducing the number of samples.

3. Discussion

This section elaborates on the main result and the considered weighted programs. Comparison to previous work is conducted in Section 3.1 and a discussion of the novel incoherence parameters is provided in Section 3.2.

3.1 Related Work

Many authors have considered how to efficiently include prior information into matrix reconstruction problems. The papers by Aravkin et al. 2014; Zhang et al. 2019, 2020 focus on applications to seismology and numerical aspects of the problem, adopting an approach that spurred from program (5) first proposed by Aravkin et al. 2014. A distinct approach is considered by Chen et al. 2015, 2014; Eftekhari et al. 2018a, where the prior information is used to bias the sampling scheme according to the array's leverage scores. Therein, the authors show that $\mathcal{O}(n_1 r\log^2(n_1))$ revealed entries provide exact (noiseless) completion of a rank r matrix with more lenient dependency on the standard incoherence parameter μ_0. Most relevant to the context of this paper, are the results by Yi et al. 2013; Bayat and Daei 2020; Chiang et al. 2015; Eftekhari et al. 2018b; Chen 2015; Xu et al. 2013; Jain and Dhillon 2013 which provide theoretical analysis related to program (5) or program (4, T_2) in Theorem 7.

The work by Eftekhari et al. 2018b applies directly to program (5) in the matrix completion case and general matrix sensing scenario. In the matrix completion setting the authors obtain sample complexity $|\Omega| \sim \mu_0 n_1 r\log(n_1)$ under inexact prior information, thereby reducing the number of samples by a logarithmic factor in contrast to unbiased nuclear norm minimization. The sample complexity and error bounds produced therein depend on the PABS (13), imposing a fidelity requisite on T for the results to be applicable. The results in this paper are not directly comparable. Arguably, the result most related in this work is provided in Theorem 10, where program (4, T_4) requires $|\Omega| \sim \mu_0(n_1 - \mu_2 r)r\log^3(n_1)$ samples. In contrast to the work by Eftekhari et al. 2018b, the result here does not require PABS-based prior knowledge or requisites for applicability and the resulting error bounds are more lenient in terms of the dependence on $\sin(\theta_u)$ and $\sin(\theta_v)$. The authors Bayat and Daei 2020 consider a more flexible program than the original approach, where four weight choices are allowed. Their results only apply to the matrix sensing scenario, where the authors demonstrate an $\mathcal{O}(nr)$ sample complexity.

The remaining citations (Yi et al. 2013; Chen 2015; Xu et al. 2013; Jain and Dhillon 2013; Abernethy et al. 2009) consider approaches arguably similar in nature to program (4, T_2) with $\omega = 0$ and exact prior information, while Chiang et al. 2015 allows for noisy side information. Among these, the smallest sample complexity is $\mathcal{O}(\mu_0 r^2 \log(n_1)\log(r))$ provable in the setting of exact prior knowledge (in the analogous scenario where $\rho = r$ for simplicity). In this context, Theorem 7 presented here with small weights allows for accurate estimation from $\mathcal{O}(\mu_0^2 r^2 \log^3(n_1))$ sampled entries. This sampling condition is slightly worse than what is derived by Yi et al. 2013; Chen 2015. However, the approach considered here is a more flexible methodology with weight selection that allows for improved reconstruction output (see Section 4 for numerical behaviour based on weight selection). Furthermore, Theorem 7 applies to cases with inexact prior knowledge and derives error bounds that provide insight when inaccurate estimates $\tilde{U}$ and $\tilde{V}$ are incorporated in different ways.

3.2 Incoherence Parameter μ_1

This section discusses the parameter μ_1 defined in (3), comparing it to the standard and joint incoherence parameters from the literature. To the best of the author's knowledge, this definition of incoherence has not appeared in the matrix completion literature. Previous optimal results have sample complexity requiring only linear dependence on the standard parameter μ_0. The results here also depend on this parameter, but additionally introduce μ_1 with sub-linear and linear dependence.

Arguably, μ_1 is reminiscent of the joint incoherence (or strong incoherence) condition introduced by Recht 2011; Gross 2011. The joint incoherence parameter will be denoted here as $\tilde{\mu}_1$. Given $r \leq n_2$ and recalling the SVD of D, the joint incoherence parameter of D is defined as the smallest $\tilde{\mu}_1 > 0$ such that

$$\max_{(k,\ell)\in[n_1]\times[n_2]} \left| \sum_{j=1}^{r} U_{kj}\bar{V}_{\ell j} \right| \leq \sqrt{\frac{\tilde{\mu}_1 r}{n_1 n_2}}.$$

In particular, μ_1 in definition (3) also depends jointly on the right and left singular vectors. Furthermore, $\mu_1 \leq \mu_0^2 r$ which is also a tight upper bound for the joint incoherence parameter (Chen 2015). However, it is important to note that Recht 2011 and other authors derived sample complexity $m \sim \tilde{\mu}_1 n_1 r \log^2(n_1)$. In contrast, the work here requires $m \sim \sqrt{\mu_1} n_1 r \log^2(n_1)$ in Theorem 2 and $m \sim \mu_1 r \log^3(n_1)$ in Theorem 6. Though it is difficult to provide a fair comparison, this section will argue that the results here impose relatively lenient incoherent conditions in a "joint" sense.

The author Chen 2015 discusses the exorbitant nature of the joint incoherence parameter since it intuitively requires the rows of U^r and V^r to be unaligned, a requisite with no reasonable explanation. In the current work, all results would also hold if μ_1 were defined as the smallest number such that

$$\max_{1\leq k\leq n_1} \left(\sum_{j=1}^{r} |U_{kj}|^4 \right)^{1/4} \max_{1\leq \ell\leq n_2} \left(\sum_{j=1}^{r} |V_{\ell j}|^4 \right)^{1/4} \leq \sqrt{\frac{\mu_1 r}{n_1 n_2}}. \tag{20}$$

This alternative definition alleviates the joint nature of the original definition. It is now arguable that this condition requires the ℓ_4 norms of the rows of U^r and V^r to both be small, thereby imposing an additional "non-spikiness" condition analogous to the requisite of a small μ_0 parameter with respect to the ℓ_2 norms. However, in contrast the adopted definition (3), the incoherence parameter given by (20) is pessimistic (see the second example below) and for this reason the original definition is kept.

Notice that Theorem 2 only has sub-linear dependence on the introduced parameter $\sqrt{\mu_1}$. This observation is crucial to properly compare some of the work here with previous incoherence optimal results. To elaborate, specific data matrices are produced to compute μ_0, $\sqrt{\mu_1}$, and $\tilde{\mu}_1$.

- **Case $\mu_0 = \sqrt{\mu_1}$:** consider the random orthogonal model and the incoherent basis model by Candès and Recht 2009. Therein, the authors show that a rank r matrix $M = U\Sigma V^\top$ generated from the random orthogonal model obeys $\max_{k,\ell} |U_{k\ell}|^2 \leq 10\log(n_1)/n_1$ and $\max_{k,\ell} |V_{k\ell}|^2 \leq 10\log(n_2)/n_2$ with high probability. From this, it is easy to see that $\mu_0, \sqrt{\mu_1} \sim \log(n_1)$. A similar conclusion holds trivially for the incoherent basis model and any singular vectors obeying the size property 1.12 in the same reference.

  Joint incoherence: in this example, $\tilde{\mu}_1 \sim \log^2(n_1)$ (see Candès and Recht 2009).

- **Case $\mu_0 > \sqrt{\mu_1}$:** let $r < n_2$ and $M = UV^\top$ where the columns of U and V consist of any r columns of I_{n_1} (the $n_1 \times n_1$ identity matrix) and any r columns of $\mathcal{F} : \mathbb{C}^{n_2} \mapsto \mathbb{C}^{n_2}$ (the 1D Fourier transform) respectively. Then the r-incoherence parameters satisfy $\mu_0 = n_1/r$ and $\sqrt{\mu_1} = \sqrt{n_1/r}$. Note that definition (20) would instead obtain $\sqrt{\mu_1} = \sqrt{n_1/\sqrt{r}}$ in this example, which demonstrates the improvement gained by the chosen definition (3).

  Joint incoherence: here $\tilde{\mu}_1 \sim n_1/r$.

- **Case $\mu_0 < \sqrt{\mu_1}$:** let $M = UV^\top$ where the columns of U and V consist of any r columns of I_{n_1} and any r columns of I_{n_2} respectively. This example gives the worst case r-incoherence parameters $\mu_0 = n_1/r$ and $\sqrt{\mu_1} = \sqrt{n_1 n_2/r}$. In general, using (20) shows that $\sqrt{\mu_1} \leq \mu_0\sqrt{r}$, which is sharp according to this example when $n_1 = n_2$.

  Joint incoherence: $\tilde{\mu}_1 \sim n_1 n_2/r$.

From these examples, it is clear that there is no strict relationship between μ_0 and $\sqrt{\mu_1}$. Moreover, typical data matrices of interest from the literature seem to largely lie in the regime where $\sqrt{\mu_1} \leq \mu_0$. In these cases, Theorem 2 intuitively reduces to solely depend on μ_0.

Moreover, the examples reveal that $\mu_1 \sim \tilde{\mu}_1$ holds intuitively, though a proof of such a statement is not provided in this work. The lenient joint incoherence conditions of this paper are due to the sub-linear dependence $\sqrt{\mu_1}$ in Theorem 2, which is most comparable to the work of Recht 2011. Notice that $\sqrt{\mu_1} \leq \tilde{\mu}_1$ holds in the examples above. However, Theorem 6 does require linear dependence on μ_1 (but removes the linear dependence on n_1). Since μ_1 may be as large as $r\mu_0^2$, this may lead to the requisite $m \sim r^2\mu_0^2\log^3(n_1)$ which still offers a severe reduction in the number of observed entries comparable to the result in Theorem 7.

4. Numerical Experiments

This section numerically explores programs (4, T_1) and (4, T_2), comparing them to the original weighted nuclear norm minimization program (5). The goal of this section is to numerically validate the error bounds in Section 2.1. The experiments will reveal the practicality of the derived analysis, agreeing with the theoretical observation that weighted programs incorporating subspaces that require less samples will in general exhibit increasing sensitivity of inaccurate prior information.

The setup of Eftekhari et al. 2018b is adopted to generate a data matrix and subspace information. Let $D = U^r\Sigma^r V^{r\top} \in \mathbb{R}^{n_1\times n_2}$, where $U^r \in \mathbb{R}^{n_1\times r}$ and $V^r \in \mathbb{R}^{n_2\times r}$ are constructed by orthogonalizing the columns of a standard random Gaussian matrix with r columns and normalizing so that $\|D\|_F = 1$. To obtain prior knowledge, a perturbed matrix is generated $\tilde{D} = D + N$ where the entries of $N \in \mathbb{R}^{n_1\times n_2}$ are i.i.d. Gaussian random variables with variance σ^2 that will be toggled to select a desired PABS. Then $\tilde{U} \in \mathbb{R}^{n_1\times r}$ and $\tilde{V} \in \mathbb{R}^{n_2\times r}$ are the leading r left and right singular vectors of $\tilde{D}$. The dimensions are set to $n_1 = n_2 = 500$ and $r = 50$. The set of observed matrix entries is selected uniformly at random from all subsets of the same cardinality $|\Omega| = \lambda(n_1n_2)$, where $\lambda \in [0, 1]$ will be varied to specify a desired sampling percentage. In each experiment, D, N and Ω are generated independently and programs (4) and (5) are solved with $\omega = \omega_1\omega_2$ varying in (0,1] (setting $\omega_1 = \omega_2$). The plots below present the average relative errors of 100 independent trials via trustworthy and relatively inaccurate subspace estimates.

Programs (4) and (5) are solved using the LR-BPDN implementation introduced by Aravkin et al. 2014, which combines the Pareto curve methodology (van den Berg and Friedlander 2009) with a matrix factorization approach. With $\omega > 0$, (26) is solved in lieu of (4), which is an equivalent formulation that trades off the objective function with a modified projection operator in the constraint (see A.3). This allows LR-BPDN to be directly applicable to (26), with output $\tilde{D}^\omega = \omega\mathcal{P}_T(D^\omega) + \mathcal{P}_{T^\perp}(D^\omega)$ which gives the desired solution as $D^\omega = \omega^{-1}\mathcal{P}_T(\tilde{D}^\omega) + \mathcal{P}_{T^\perp}(\tilde{D}^\omega)$. An analogous trick is used to solve (5) when $\omega_1\omega_2 > 0$.

Varying weights: noiseless numerical results with varying weights $\omega, \omega_1\omega_2 \in (0, 1]$ are shown in Figure 1. Two plots are provided, corresponding to reliable prior information with $|\Omega|/n_1n_2 = .01$ (left plot) and less accurate prior knowledge with $|\Omega|/n_1n_2 = .15$ (right plot). In these plots, the variance of N is chosen to provide PABS $\sin(\theta_u), \sin(\theta_v) \approx .1$ and $\sin(\theta_u), \sin(\theta_v) \approx .2$ respectively.

In the case of good subspace estimates (left plot) it is clear that program (4, T_1) greatly outperforms the other approaches, obtaining a relative error $\approx .1$ with only 1% of observed entries. However, this approach is demonstrated to be relatively sensitive when less accurate prior information is supplied. With relatively inaccurate prior information, the original weighted program (5) exhibits the best reconstruction error in the right plot of Figure 1. The numerical behavior illustrated in Figure 1 agrees with the derived error bounds and discussion provided in Section 2.1, where Theorem 10 provides some insight for the robust behavior of program (5) at the cost of higher sample complexity.

Varying sampling percentage: noiseless numerical results with varying percentages of observed noiseless entries are shown in Figure 2. Applying the choice of weights from Figure 1 that give the smallest reconstruction error for each program, two plots are shown

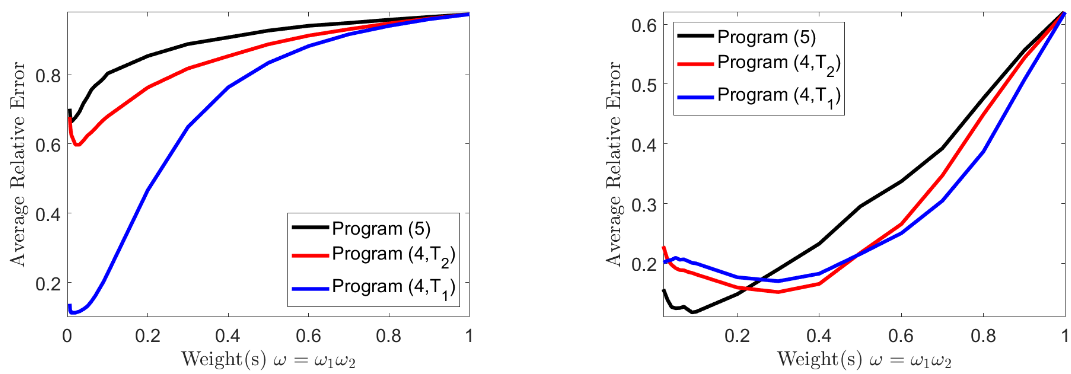

Figure 1: Plots of weight choice versus average relative error of matrix reconstruction via noiseless weighted nuclear norm minimization programs. The left plot applies reliable prior information with $\sin(\theta_u), \sin(\theta_v) \approx .1$ and %1 percent of observed matrix entries. The right plot was obtained with $\sin(\theta_u), \sin(\theta_v) \approx .2$, where %15 of the entries are observed.

with reliable subspaces (left plot) and less accurate prior knowledge (right plot). In the case of accurate estimate subspaces, program (4, T_1) obtains the smallest relative errors in all shown sampling percentages. However, observe that programs (4, T_2) and (5) obtain comparable relative errors from as little as 8% of observed entries. As in Figure 1, in the case of less accurate subspaces, the original weighted program exhibits the most flexibility toward untrustworthy prior information. Therefore, Figure 2 also reflects the trade-off behavior of sample complexity vs sensitivity to inaccurate prior information that agrees with the theoretical conclusions of Section 2.1.

5. Conclusion

In this paper, a family of weighted matrix completion programs is proposed as a means to incorporate prior knowledge into the matrix completion problem. The main result establishes the nearly optimal sampling rate $\mathcal{O}(M_1 r \log^3(n))$ by which an $n \times n$ rank r matrix can be accurately approximated when a subspace of rank r matrices is enforced (where M_1 captures dimensional and incoherence-based properties of the subspace). The analysis allows for robust matrix approximation in the case of inexact subspaces, measurement noise, and full rank data matrices. The work provides novel intuition for previous methodologies in the literature, while introducing novel approaches. Finally, the results and numerical experiments showcase an insightful trade-off caused by incorporating distinct subspaces. In general, it is observed that subspaces requiring less samples for exact matrix completion are more susceptible to inaccurate prior information.

Several important limitations and potential improvements need to be elaborated and explored as future work. The subspace incoherence parameter M_1 may extract the parameter μ_1, which is distinct in comparison to results that only require the standard incoherence condition. Modified analysis that only depends on the standard parameter μ_0 would be

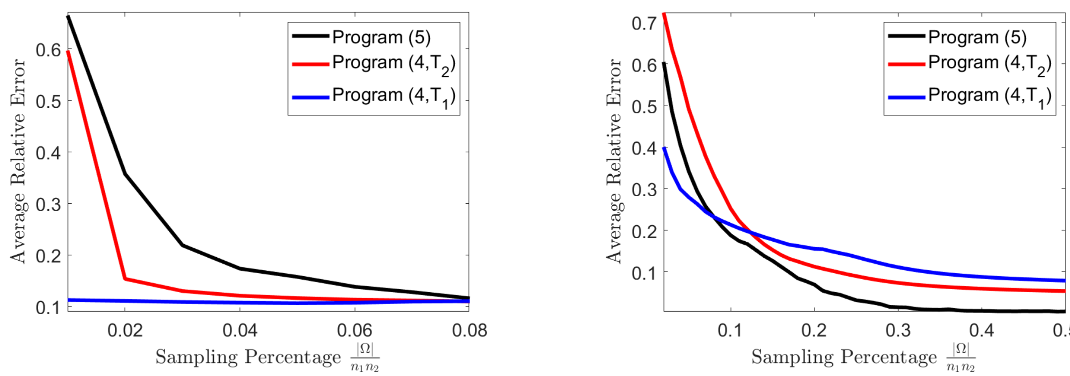

Figure 2: Plots of sampling percentage versus average relative error of matrix reconstruction via noiseless weighted nuclear norm minimization programs. The left plot applies reliable prior information with $\sin(\theta_u), \sin(\theta_v) \approx .1$ and weight choices of $\omega_1\omega_2 = .01$ for program 5, $\omega = .02$ for program (4, T_2), and $\omega = .01$ for program (4, T_1). The right plot was obtained with $\sin(\theta_u), \sin(\theta_v) \approx .2$ and weight choices of $\omega_1\omega_2 = .06$ for program 5, $\omega = .3$ for program (4, T_2), and $\omega = .2$ for program (4, T_1).

most informative and align best with the existing literature. Furthermore, the main theoretical result does not seem to provide any insight for weight selection according to the user's confidence in the estimate subspace or any other variables. An error bound that specifies the optimal weight choice in different scenarios would be of great value to practitioners. As future work, it would be of interest to adapt the proof strategy and tools of this work to further comprehend how to set the program parameters of the proposed method and previously considered weighted approaches.

Another avenue to extend this work is to consider numerically efficient alternatives to the nuclear norm approach presented here. Strong prior information may be sufficient to alleviate sample complexity, potentially rendering the low rank assumption unnecessary for matrix completion. For example, a least squares-based weighted program might arguably produce comparable estimates while reducing computational complexity by avoiding rank penalization terms. The tools developed here may be useful to analyze such approaches, and help better understand the independent roles that prior information and data rank play in the sample complexity.

Appendix A. Proof of the Main Results

This section provides proofs for the main results, only stating the required lemmas. These lemmas will be proven in Appendix C.

A.1 Dual Certificate

The first lemma establishes dual certificate conditions relative to T for recovery error bounds. It is stated in general form, applicable to any linear operator. For the statement, it is important to notice that for any $X \in \mathbb{C}^{n_1 \times n_2}$ there exists a $G \in T \cap S_{op}$ such that $\|\mathcal{P}_T(D)\|_* = \langle D, G \rangle$ by characterization of the nuclear norm via its dual norm (the operator norm). Furthermore, for $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \mapsto \mathbb{C}^m$ define

$$\|\mathcal{A}\|_{F \mapsto 2} := \max_{X \in S} \|\mathcal{A}(X)\|_2.$$

Lemma 11 *Let $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \mapsto \mathbb{C}^m$ be a linear operator, and*

$$\inf_{X \in T \cap S} \|\mathcal{A}(X)\|_2 \geq \beta_1 > 0, \qquad \|\mathcal{A}\mathcal{P}_{T^\perp}\|_{F \mapsto 2} \leq \beta_2. \tag{21}$$

Given $D \in \mathbb{C}^{n_1 \times n_2}$ with $\|\mathcal{P}_T(D)\|_* = \langle D, G \rangle$ for some $G \in T \cap S_{op}$, assume that there exist $Y, Z \in \mathbb{C}^{n_1 \times n_2}$ with $Y = \mathcal{A}^*\mathcal{A}(Z)$ satisfying

$$\|\mathcal{P}_T(Y) - G\|_F \leq \beta_3, \qquad \|\mathcal{P}_{T^\perp}(Y)\| \leq \beta_4, \qquad \|\mathcal{A}(Z)\|_2 \leq \beta_5 \tag{22}$$

and $\frac{\beta_2 \beta_3}{\beta_1} + \beta_4 < 1$. Let $d \in \mathbb{C}^m$ with $\|d\|_2 \leq \eta$ and

$$D^\sharp := \operatorname*{argmin}_{X \in \mathbb{C}^{n_1 \times n_2}} \|X\|_* \quad s.t. \quad \|\mathcal{A}(X) - \mathcal{A}(D) - d\|_2 \leq \eta.$$

Then

$$\|D - D^\sharp\|_F \leq C_1 \|\mathcal{P}_{T^\perp}(D)\|_* + C_2 \eta,$$

where C_1, C_2 depend on the β_k's.

The proof of this lemma is postponed until C.1, where the dependence of C_1 and C_2 on the β_k's is specified. The main result will be obtained by establishing (21) and (22) for the matrix completion sampling operator P_Ω. With this in mind, the proof of Theorem 4 is provided, from which the remaining theorems are corollaries (see Appendix B). The proof considers the sampling with replacement model, discussed in detail in the next section.

A.2 Sampling Model

As in previous work, the work load will be simplified by considering the uniform sampling with replacement model. In other words, let $\tilde{\Omega}$ be generated by choosing m entries independently and uniformly at random from $[n_1] \times [n_2]$ (this allows $\tilde{\Omega}$ to have repeated entries). Recht 2011 and Gross 2011 show that any upper bound on the probability of failure for exact (noiseless) matrix completion via $\tilde{\Omega}$ is also valid for uniform sampling without replacement. This strategy will apply in the scenario of this work, proceeding in a different manner to include noisy observations, full rank matrices, and inexact subspace estimates.

With $\tilde{\Omega}$ generated as above, let $\Omega \subseteq \tilde{\Omega}$ consist of the $|\Omega| \leq m$ distinct samples in $\tilde{\Omega}$ and define the normalized operators

$$\tilde{\mathcal{A}} = \sqrt{\frac{n_1 n_2}{m}} P_{\tilde{\Omega}} \quad \text{and} \quad \mathcal{A} = \sqrt{\frac{n_1 n_2}{m}} P_\Omega,$$

which extract and scale an input matrix's values in the entries specified by the subset in the subscript. As shown by Gross 2011 (Section II), the distribution of Ω is the same as the distribution of sampling $\left|\Omega\right|$ entries uniformly without replacement. Therefore, assume $\Omega \subseteq \tilde{\Omega}$ WLOG, so that generating $\tilde{\Omega}$ with m entries will generate Ω uniformly at random and any lower bound requisite for $|\Omega|$ will also be satisfied by m (since $m \geq |\Omega|$).

Notice that for any $X \in \mathbb{C}^{n_1 \times n_2}$ and $(k, \ell) \in [n_1] \times [n_2]$

$$(P_\Omega^* P_\Omega(X))_{k\ell} = \begin{cases} X_{k\ell} & \text{if } (k,\ell) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

and

$$\left(P_{\tilde{\Omega}}^* P_{\tilde{\Omega}}(X)\right)_{k\ell} = \begin{cases} \text{mult}_{\tilde{\Omega}}(k,\ell) X_{k\ell} & \text{if } (k,\ell) \in \tilde{\Omega} \\ 0 & \text{otherwise} \end{cases}$$

where $\text{mult}_{\tilde{\Omega}}(k,\ell) \in \mathbb{N}$ is the multiplicity of $(k,\ell) \in \tilde{\Omega}$. Let $\text{mult}_{\tilde{\Omega}}(k,\ell) \leq \tau$ for all $(k,\ell) \in \tilde{\Omega}$. The quantity τ will be bounded using Proposition 3.3 by Recht 2011 with parameter $\beta = 3/2$ therein. For $n_1 \geq 9$, the proposition gives $\tau \leq 4\log(n_1)$ with probability exceeding $1 - n_1^{-1}$. Then for any $X \in \mathbb{C}^{n_1 \times n_2}$

$$\|\tilde{\mathcal{A}}(X)\|_2 \leq \sqrt{\tau}\|\mathcal{A}(X)\|_2 \leq 2\sqrt{\log(n_1)}\|\mathcal{A}(X)\|_2. \tag{23}$$

The remainder will operate in this scenario so that (23) holds with high probability. Therefore, using the lower bound

$$\inf_{X \in T \cap S} \|\tilde{\mathcal{A}}(X)\|_2 \geq \tilde{\beta}_1$$

for some $\tilde{\beta}_1 > 0$, allows to choose parameter $\beta_1 = \frac{\tilde{\beta}_1}{2\sqrt{\log(n_1)}} > 0$ from Lemma 11. The following lemma is crucial for this purpose, and will also be used to compute parameters β_3 and β_5 with an appropriate choice of dual certificate.

Lemma 12 *Let $T \subset \mathbb{C}^{n_1 \times n_2}$ be a subspace with subspace joint incoherence parameter M_1 and ρ defined as in (8). With $\tilde{\mathcal{A}}$ as above and $0 < \delta \leq \frac{1}{4}$, if*

$$\sqrt{m} \geq \frac{C\sqrt{M_1 \rho}\log^{3/2}(n_1 + n_2)}{\delta}$$

where $C > 0$ is an absolute constant, then

$$\sup_{X \in T \cap S} \left|\langle(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}} - \mathcal{I})(X), X\rangle\right| < 2\delta,$$

with probability exceeding

$$1 - \exp\left(-\frac{6m\delta^2}{19M_1\rho}\right).$$

The proof can be found in C.2, which generalizes the approach of Rauhut 2008; Rudelson and Vershynin 2008 (Rudelson-Vershynin Lemma via Dudley's inequality) as done by Liu 2011. The following lemma by Chen 2015 is needed to compute β_4.

Lemma 13 *Suppose $Z \in \mathbb{C}^{n_1 \times n_2}$ is a fixed matrix. Then with probability greater than $1 - \frac{1}{n_1+n_2}$*

$$\|\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}(Z) - Z\| \leq \frac{4n_1n_2\log(n_1+n_2)}{3m}\|Z\|_\infty + 2\sqrt{\frac{2n_1n_2\log(n_1+n_2)}{m}}\|Z\|_{\infty,2}.$$

The concentration inequality uses the norm

$$\|Z\|_{\infty,2} := \max\left\{\max_k \sqrt{\sum_j |Z_{kj}|^2}, \max_\ell \sqrt{\sum_j |Z_{j\ell}|^2}\right\} \tag{24}$$

which is the maximum of the row and column norms of Z. The result appears as Lemma 2 in the work of Chen 2015, the proof is adapted here to the sampling with replacement model (see C.3 for the proof). The parameters in (21) and (22) may now be computed to establish the weighted matrix completion error bounds.

A.3 Proof of Theorem 4: Weighted Matrix Completion

With respect to the normalized operators, the output of (4) is equivalently given as

$$D^\omega := \underset{X\in\mathbb{C}^{n_1\times n_2}}{\text{argmin}} \left\|\omega\mathcal{P}_T(X) + \mathcal{P}_{T^\perp}(X)\right\|_* \text{ s.t. } \left\|\mathcal{A}(D) + \sqrt{\frac{n_1n_2}{m}}d - \mathcal{A}(X)\right\|_2 \leq \sqrt{\frac{n_1n_2}{m}}\eta.$$

Further note that

$$\begin{aligned} &\omega\mathcal{P}_T(D^\omega) + \mathcal{P}_{T^\perp}(D^\omega) = \\ &\underset{X\in\mathbb{C}^{n_1\times n_2}}{\text{argmin}} \|X\|_* \text{ s.t. } \left\|\mathcal{A}(D) + \sqrt{\frac{n_1n_2}{m}}d - \mathcal{A}\left(\frac{1}{\omega}\mathcal{P}_T(X) + \mathcal{P}_{T^\perp}(X)\right)\right\|_2 \leq \sqrt{\frac{n_1n_2}{m}}\eta, \end{aligned} \tag{25}$$

where equality holds since $\omega\mathcal{P}_T(\circ)+\mathcal{P}_{T^\perp}(\circ)$ is invertible when $\omega > 0$, with inverse operator $\omega^{-1}\mathcal{P}_T(\circ) + \mathcal{P}_{T^\perp}(\circ)$. Let

$$\tilde{\mathcal{P}}_\omega(\circ) = \mathcal{P}_T(\circ) + \omega\mathcal{P}_{T^\perp}(\circ).$$

Multiplying both sides of the constraint in (25) by ω gives

$$\begin{aligned} &\omega\mathcal{P}_T(D^\omega) + \mathcal{P}_{T^\perp}(D^\omega) = \\ &\underset{X\in\mathbb{C}^{n_1\times n_2}}{\text{argmin}} \|X\|_* \text{ s.t. } \left\|\omega\mathcal{A}(D) + \omega\sqrt{\frac{n_1n_2}{m}}d - \mathcal{A}\left(\tilde{\mathcal{P}}_\omega(X)\right)\right\|_2 \leq \omega\sqrt{\frac{n_1n_2}{m}}\eta \\ &= \underset{X\in\mathbb{C}^{n_1\times n_2}}{\text{argmin}} \|X\|_* \text{ s.t. } \left\|\mathcal{A}\left(\tilde{\mathcal{P}}_\omega(D_T)\right) + \omega\sqrt{\frac{n_1n_2}{m}}d - \mathcal{A}\left(\tilde{\mathcal{P}}_\omega(X)\right)\right\|_2 \leq \omega\sqrt{\frac{n_1n_2}{m}}\eta, \end{aligned} \tag{26}$$

were the last line defines $D_T := \omega\mathcal{P}_T(D) + \mathcal{P}_{T^\perp}(D)$.

The dual certificate of Lemma 11 will be produced with respect to program (26) with sampling operator $\mathcal{A}_T := \mathcal{A}\tilde{\mathcal{P}}_\omega$ to obtain an upper bound for

$$\left\|D_T - \omega\mathcal{P}_T(D^\omega) + \mathcal{P}_{T^\perp}(D^\omega)\right\|_F,$$

which will in turn will give an appropriate bound for $\|D - D^\omega\|_F$.
Proof [Proof of Theorem 4]
Using the notation above, notice that

$$\inf_{X \in T \cap S} \|\mathcal{A}_T(X)\|_2 = \inf_{X \in T \cap S} \|\mathcal{A}(X)\|_2. \tag{27}$$

The parameters in (21) and (22) from Theorem 11 with respect to $\mathcal{A}_T$, D_T, $\tilde{U}$ and $\tilde{V}$ will now be bounded. Let $G \in T \cap S_{op}$ be such that $\|\mathcal{P}_T(D_T)\|_* = \langle D_T, G\rangle$. For the dual certificate, choose

$$Y := \mathcal{A}_T^* \mathcal{A}_T(Z),$$

where

$$Z := \tilde{\mathcal{P}}_\omega^{-1} \left(P_{\tilde{\Omega}}^* P_{\tilde{\Omega}}(G) \right).$$

Notice that

$$Y := \mathcal{A}_T^* \mathcal{A} \left(P_{\tilde{\Omega}}^* P_{\tilde{\Omega}}(G) \right).$$

After bounding the β_k's, the proof will finish by applying Lemma 11.

- Parameter β_1: using (27) and (23) will give $\beta_1 = (2\sqrt{2\log(n_1)})^{-1}$. To show this, Theorem 12 will be applied in two different ways below with $\delta \leq 1/4$ so that for any $X \in T \cap S$
$$\left| \langle (\tilde{\mathcal{A}}^* \tilde{\mathcal{A}} - \mathcal{I})(X), X \rangle \right| < 2\delta \leq \frac{1}{2} \tag{28}$$
and consequently $\|\tilde{\mathcal{A}}(X)\|_2 \geq \sqrt{1 - 2\delta} \geq 1/\sqrt{2}$ which gives $\beta_1 \geq (2\sqrt{2\log(n_1)})^{-1}$ with high probability.

- Parameter β_2: this parameter is $\|\mathcal{A}_T \mathcal{P}_{T^\perp}\|_{F\to 2} = \omega \|\mathcal{A}\mathcal{P}_{T^\perp}\|_{F \to 2}$. With the chosen dual certificate, use the following calculation
$$\omega \big\|\mathcal{A}\mathcal{P}_{T^\perp}\big\|_{F\to 2} \leq \omega \sqrt{\frac{n_1 n_2}{m}} \|P_\Omega\|_{F\to 2} \|\mathcal{P}_{T^\perp}\|_{F \to F} \leq \omega \sqrt{\frac{n_1 n_2}{m}} := \beta_2,$$
where $\|P_\Omega\|_{F\to 2} \leq 1$ holds since Ω has no repeated entries.

- Parameter β_3: this parameter can be computed using (28). Note that
$$\mathcal{P}_T(Y) = \mathcal{P}_T \circ \mathcal{A}^* \mathcal{A} \left(P_{\tilde{\Omega}}^* P_{\tilde{\Omega}}(G) \right) = \mathcal{P}_T \circ \tilde{\mathcal{A}}^* \tilde{\mathcal{A}} (G).$$
Therefore,
$$\|\mathcal{P}_T(Y) - G\|_F = \|\mathcal{P}_T \circ (\tilde{\mathcal{A}}^* \tilde{\mathcal{A}} - \mathcal{I}) \circ \mathcal{P}_T(G)\|_F \leq 2\delta\sqrt{\rho} := \beta_3,$$
since $\|G\| = 1$ gives $\|G\|_F \leq \sqrt{\operatorname{rank}(G)} \leq \sqrt{\rho}$ and
$$\sup_{X \in T \cap S} \left| \langle (\tilde{\mathcal{A}}^* \tilde{\mathcal{A}} - \mathcal{I})(X), X \rangle \right| = \|\mathcal{P}_T \circ (\tilde{\mathcal{A}}^* \tilde{\mathcal{A}} - \mathcal{I}) \circ \mathcal{P}_T\|_{F \mapsto F} \leq 2\delta.$$
Later, δ will be chosen in two different ways (always satisfying $\delta \leq 1/4$) which will change the value of β_3 in each scenario.

- Parameter β_5: under the scenario $\text{mult}(k,\ell) \leq \tau \leq 4\log(n_1)$ (see discussion in A.2), it can be shown that $\beta_5 \leq \sqrt{6\rho\log(n_1)}$ as follows

$$\begin{aligned}\left\|\mathcal{A}(P_{\tilde{\Omega}}^* P_{\tilde{\Omega}}(G))\right\|_2 &:= \sqrt{\frac{n_1 n_2}{m}}\left\|P_\Omega\left(P_{\tilde{\Omega}}^* P_{\tilde{\Omega}}(G)\right)\right\|_2 = \sqrt{\frac{n_1 n_2}{m}}\left\|P_{\tilde{\Omega}}^* P_{\tilde{\Omega}}(G)\right\|_F \\ &\leq \sqrt{\frac{\tau n_1 n_2}{m}}\left\|P_{\tilde{\Omega}}(\tilde{U}\tilde{\Sigma}\tilde{V}^\top)\right\|_2 := \sqrt{\tau}\|\tilde{\mathcal{A}}(G)\|_2 \leq \sqrt{6\log(n_1)}\|G\|_F.\end{aligned}$$

  The last inequality follows from (28) with $2\delta \leq 1/2$, since $G \in T$.

The remaining parameter β_4 is bounded by considering distinct ranges for ω and choices for $\delta \leq 1/4$ (which will also determine β_3).

Case $\omega \leq \frac{\sqrt{M_1}\log(n_1+n_2)}{\sqrt{n_1 n_2}}$: in this scenario, apply Lemma 12 with $\delta = 1/4$, which holds if

$$\sqrt{m} \geq 4C\sqrt{M_1\rho}\log^{3/2}(n_1+n_2). \tag{29}$$

With appropriate C, the probability of success exceeds $1-(n_1+n_2)^{-1}$. Notice that this case gives $\beta_3 = \sqrt{\rho}/2$.

- Parameter β_4: to bound $\left\|\mathcal{P}_{T^\perp}(Y)\right\|$, notice that

$$\begin{aligned}\left\|\mathcal{P}_{T^\perp}(Y)\right\| &= \omega\left\|\mathcal{P}_{T^\perp}\circ\mathcal{A}^*\mathcal{A}\left(P_{\tilde{\Omega}}^* P_{\tilde{\Omega}}(G)\right)\right\| = \omega\left\|\mathcal{P}_{T^\perp}\left(\mathcal{A}^*\mathcal{A}\left(P_{\tilde{\Omega}}^* P_{\tilde{\Omega}}(G)\right) - G\right)\right\| \\ &\leq \omega\left\|\mathcal{A}^*\mathcal{A}\left(P_{\tilde{\Omega}}^* P_{\tilde{\Omega}}(G)\right) - G\right\| = \omega\left\|\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}\left(G\right) - G\right\|,\end{aligned}$$

  where the inequality holds since $\mathcal{P}_{T^\perp}$ is an orthogonal projection. Using the bound above and Lemma 13 gives

$$\left\|\mathcal{P}_{T^\perp}(Y)\right\| \leq \frac{4\omega n_1 n_2\log(n_1+n_2)}{3m}\|G\|_\infty + 2\omega\sqrt{\frac{2n_1 n_2\log(n_1+n_2)}{m}}\|G\|_{\infty,2}$$

  with probability at least $1-(n_1+n_2)^{-1}$. Using the subspace incoherence condition, it is clear that

$$\|G\|_{\infty,2} \leq \sqrt{\frac{M_0\rho}{n_2}} \quad \text{and} \quad \|G\|_\infty \leq \|G\|_F\sqrt{\frac{M_1\rho}{n_1 n_2}} \leq \rho\sqrt{\frac{M_1}{n_1 n_2}}. \tag{30}$$

  Using (29) and (30) gives

$$\left\|\mathcal{P}_{T^\perp}(Y)\right\| \leq \frac{\omega\sqrt{n_1 n_2}}{12C^2\sqrt{M_1}\log^2(n_1+n_2)} + \frac{\omega\sqrt{M_0 n_1}}{\sqrt{2}C\sqrt{M_1}\log(n_1+n_2)} := \beta_4$$

  and

$$\frac{\beta_2\beta_3}{\beta_1} = \frac{\omega\sqrt{2n_1 n_2\rho\log(n_1)}}{\sqrt{m}} \leq \frac{\omega\sqrt{n_1 n_2}}{2C\sqrt{2M_1}\log(n_1+n_2)}.$$

  Note that $M_0 \leq n_2$, and therefore $\frac{\beta_2\beta_3}{\beta_1} + \beta_4 < 1$ if $\omega \leq \frac{\sqrt{M_1}\log(n_1+n_2)}{\sqrt{n_1 n_2}}$ and Theorem 11 may now be applied.

Case $\frac{1}{2\sqrt{2\rho\log(n_1+n_2)}} \leq \omega \leq 1$: apply Lemma 12 with $\delta = \sqrt{m}/8\omega n_1\sqrt{2\rho\log(n_1+n_2)}$. Using $m \leq n_1^2$ and the assumption on ω gives

$$\delta = \frac{\sqrt{m}}{8\omega n_1\sqrt{2\rho\log(n_1+n_2)}} \leq \frac{1}{8\omega\sqrt{2\rho\log(n_1+n_2)}} \leq \frac{1}{4},$$

as desired if

$$\sqrt{m} \geq C\max\{\sqrt{M_1}, M_0\}\sqrt{\rho}\log^{3/2}(n_1+n_2)\frac{8\omega n_1\sqrt{2\rho\log(n_1+n_2)}}{\sqrt{m}}. \tag{31}$$

With an appropriate choice of C, the probability of success exceeds $1-(n_1+n_2)^{-1}$. This case gives $\beta_3 = \sqrt{m}/4\omega n_1\sqrt{2\log(n_1+n_2)}$.

- Parameter β_4: as in the previous range for ω, Lemma 13 along with (30) and (31) gives that with high probability

  $$\left\|\mathcal{P}_{T^\perp}(Y)\right\| \leq \frac{1}{6\sqrt{2}C\log(n_1+n_2)} + \frac{\sqrt{\omega}}{\sqrt{C\sqrt{2}\log(n_1+n_2)}} \leq \frac{1}{4} := \beta_4,$$

  where the last inequality holds with $\omega \leq 1$ and an appropriate choice of C. Note that $\frac{\beta_2\beta_3}{\beta_1} = \frac{\sqrt{n_2}}{2\sqrt{n_1}}$, and therefore $\frac{\beta_2\beta_3}{\beta_1} + \beta_4 = 3/4$ so that Theorem 11 can be applied in this case as well.

This concludes the bounds for all β_k parameters in (21) and (22) from Theorem 11. In both considered weight ω ranges, Theorem 11 gives constants (considering only dominating terms, see proof in C.1)

$$\begin{aligned} C_1 &\sim \frac{\beta_2}{\beta_1} + 1 = \omega\sqrt{\frac{n_1 n_2\log(n_1)}{m}} + 1, \\ C_2 &\sim \frac{\beta_2\beta_5}{\beta_1} + \beta_5 = \sqrt{\log(n_1)}\left(\omega\sqrt{\frac{n_1 n_2\rho\log(n_1)}{m}} + 1\right), \end{aligned}$$

and error bound

$$\begin{aligned} &\left\|D_T - \omega\mathcal{P}_T(D^\omega) - \mathcal{P}_{T^\perp}(D^\omega)\right\|_F \leq C_1\left\|\mathcal{P}_{T^\perp}(D_T)\right\|_* + \omega\sqrt{\frac{n_1n_2}{m}}C_2\eta \\ &= C_1\left\|\mathcal{P}_{T^\perp}(D)\right\|_* + \omega\sqrt{\frac{n_1n_2}{m}}C_2\eta. \end{aligned} \tag{32}$$

The desired error term $\|D - D^\omega\|_F$ will be bounded in two different ways to introduce $f(\omega)$ in (11) as the minimum of both bounds.

Bound 1 for $\|D - D^\omega\|_F$: notice that

$$\begin{aligned} &\left\|D_T - \omega\mathcal{P}_T(D^\omega) - \mathcal{P}_{T^\perp}(D^\omega)\right\|_F^2 = \left\|\mathcal{P}_{T^\perp}(D - D^\omega) + \omega\mathcal{P}_T(D - D^\omega)\right\|_F^2 \\ &= \left\|\mathcal{P}_{T^\perp}(D - D^\omega)\right\|_F^2 + \omega^2\left\|\mathcal{P}_T(D - D^\omega)\right\|_F^2 \\ &\geq \omega^2\|D - D^\omega\|_F^2 + \omega^2\|D - D^\omega\|_F^2 = \omega^2\|D - D^\omega\|_F^2, \end{aligned}$$

so for some absolute constant

$$\|D - D^\omega\|_F \le C\left(\sqrt{\frac{n_1n_2\log(n_1)}{m}} + \frac{1}{\omega}\right)\left\|\mathcal{P}_{T^\perp}(D)\right\|_* \\ + C\sqrt{\frac{n_1n_2\log(n_1)}{m}}\left(\omega\sqrt{\frac{n_1n_2\rho\log(n_1)}{m}} + 1\right)\eta.$$

<u>Bound 2 for $\|D - D^\omega\|_F$</u>: use the derived properties of $\mathcal{A}$ to obtain

$$\begin{aligned}
&\|D - D^\omega\|_F^2 = \|\mathcal{P}_T(D - D^\omega)\|_F^2 + \left\|\mathcal{P}_{T^\perp}(D - D^\omega)\right\|_F^2 \\
&\le \frac{1}{\beta_1^2}\|\mathcal{A}\left(\mathcal{P}_T(D - D^\omega)\right)\|_2^2 + \left\|\mathcal{P}_{T^\perp}(D - D^\omega)\right\|_F^2 \\
&\le \frac{1}{\beta_1^2}\left(\|\mathcal{A}\left(D - D^\omega\right)\|_2 + \left\|\mathcal{A}\left(\mathcal{P}_{T^\perp}(D - D^\omega)\right)\right\|_2\right)^2 + \left\|\mathcal{P}_{T^\perp}(D - D^\omega)\right\|_F^2 \\
&\le \frac{1}{\beta_1^2}\left(2\sqrt{\frac{n_1n_2}{m}}\eta + \sqrt{\frac{n_1n_2}{m}}\left\|\mathcal{P}_{T^\perp}(D - D^\omega)\right\|_F\right)^2 + \left\|\mathcal{P}_{T^\perp}(D - D^\omega)\right\|_F^2,
\end{aligned}$$

where the last inequality holds by feasibility of D^ω for (4) and since $\mathcal{A} := \sqrt{\frac{n_1n_2}{m}}P_\Omega$ contains no repeated entries. The term in the final line can be bounded by (32) since

$$\begin{aligned}
&\left\|\mathcal{P}_{T^\perp}(D - D^\omega)\right\|_F^2 \le \left\|\mathcal{P}_{T^\perp}(D - D^\omega)\right\|_F^2 + \omega\|\mathcal{P}_T(D - D^\omega)\|_F^2 \\
&= \left\|\mathcal{P}_{T^\perp}(D - D^\omega) + \omega\mathcal{P}_T(D - D^\omega)\right\|_F^2 = \left\|D_T - \omega\mathcal{P}_T(D^\omega) - \mathcal{P}_{T^\perp}(D^\omega)\right\|_F^2.
\end{aligned}$$

This approach gives that for some absolute constant

$$\|D - D^\omega\|_F \le C\sqrt{\frac{n_1n_2\log(n_1)}{m}}\left(\omega\sqrt{\frac{n_1n_2\log(n_1)}{m}} + 1\right)\left\|\mathcal{P}_{T^\perp}(D)\right\|_* \\ + C\omega\frac{n_1n_2\log(n_1)}{m}\left(\omega\sqrt{\frac{n_1n_2\rho\log(n_1)}{m}} + 1\right)\eta + \sqrt{\frac{n_1n_2\log(n_1)}{m}}\eta.$$

Notice that, in contrast to bound 1 for $\|D - D^\omega\|_F$, bound 2 includes the multiplicative term $\omega\sqrt{n_1n_2\log(n_1)/m}$ and the additive term $\sqrt{n_1n_2\log(n_1)/m}$ in the noise term. Factoring out the multiplicative term and choosing the minimum defines $f(\omega)$.

To finish, bound $\left\|\mathcal{P}_{T^\perp}(D)\right\|_*$ as follows

$$\begin{aligned}
&\left\|\mathcal{P}_{T^\perp}(D)\right\|_* \le \left\|\mathcal{P}_{T^\perp}(U^\rho\Sigma^\rho V^{\rho\top})\right\|_* + \left\|\mathcal{P}_{T^\perp}(U^+\Sigma^+V^{+\top})\right\|_* \\
&\le \left\|\mathcal{P}_{T^\perp}(U^\rho\Sigma^\rho V^{\rho\top})\right\|_* + \|U^+\Sigma^+V^{+\top}\|_* = \left\|\mathcal{P}_{T^\perp}(U^\rho\Sigma^\rho V^{\rho\top})\right\|_* + \sum_{k=\rho+1}^{n_2}\sigma_k(D).
\end{aligned}$$

■

To establish Corollary 5 from the proof above, it suffices to bound $\left\|\mathcal{P}_{T^\perp}(U^\rho\Sigma^\rho V^{\rho\top})\right\|_*$ as follows:

$$\left\|\mathcal{P}_{T^\perp}\mathcal{P}_{T_\rho}(U^\rho\Sigma^\rho V^{\rho\top})\right\|_* \le \sqrt{n_2}\left\|\mathcal{P}_{T^\perp}\mathcal{P}_{T_\rho}(U^\rho\Sigma^\rho V^{\rho\top})\right\|_F \le \sqrt{n_2}\left\|\mathcal{P}_{T^\perp}\mathcal{P}_{T_\rho}\right\|_{F\to F}\|U^\rho\Sigma^\rho V^{\rho\top}\|_F.$$

Appendix B. Proof of Theorems 2, 6, 7, 8, and 10

Theorems 6, 7, 8, and 10 are all corollaries of Theorem 4 with small ω while Theorem 2 applies $\omega = 1$. To obtain these results, it is sufficient to bound the ρ-subspace incoherence parameters for the respective estimate subspaces (all with $\rho = r$) and the term $\left\|\mathcal{P}_{T^\perp}(U^r\Sigma^r V^{r\top})\right\|_*$ in (11).

For Theorem 2, notice that with $\omega = 1$ program (4) becomes the nuclear norm minimization program (1). This makes T irrelevant, allowing the choice of subspace as in Theorem 6 with $\tilde{U} = U^r$ and $\tilde{V} = V^r$ to provide $\left\|\mathcal{P}_{T^\perp}(U^r\Sigma^r V^{r\top})\right\|_* = 0$. It only remains to upper bound M_0 in (6) and M_1 (7) in this case.

- Theorem 7, program (4, T_2): applying Theorem 4 with small weights requires upper and lower bounding the subspace joint incoherence parameter M_1. Recall that any $X \in T_2 \cap S$ satisfies $X = \tilde{U}\tilde{U}^\top X\tilde{V}\tilde{V}^\top$, so X can be written as $X = \tilde{U}W^\top$ and $X = Z\tilde{V}^\top$ where range$(W) \subset$ range$(\tilde{V})$ and range$(Z) \subset$ range$(\tilde{U})$.

  To upper bound the subspace joint incoherence, write $X = \tilde{U}W^\top$ and notice that $W = \tilde{V}\alpha$ where $\alpha \in \mathbb{C}^{r\times r}$ and $\|\alpha\|_F = 1$. Then,

$$\begin{aligned}\left|(\tilde{U}W^\top)_{pq}\right| &= \left|\sum_k \tilde{U}_{pk}\overline{W}_{qk}\right| = \left|\sum_{kj} \tilde{U}_{pk}\overline{\alpha}_{jk}\overline{\tilde{V}}_{qj}\right| \\ &\le \|\alpha\|_F \left(\sum_{kj} |\tilde{U}_{pk}|^2|\tilde{V}_{qj}|^2\right)^{1/2} \le \frac{\sqrt{\mu_0(\tilde{U}\tilde{U}^\top)\mu_0(\tilde{V}\tilde{V}^\top)r}}{\sqrt{n_1n_2}}.\end{aligned}$$

  Therefore $M_1(T_2) \le \mu_0(\tilde{U}\tilde{U}^\top)\mu_0(\tilde{V}\tilde{V}^\top)r$.

  The lower bound will be obtained by a proper selection of α. Let $\tilde{p} \in [n_1]$ obtain the maximum row norm of $\tilde{U}$ and likewise $\tilde{q} \in [n_2]$ for $\tilde{V}$. Then with $\alpha_{jk} = c\tilde{U}_{\tilde{p}k}\overline{\tilde{V}}_{\tilde{q}j}$ where c is a normalization constant (so that $\|\alpha\|_F = 1$) gives

$$\left|(\tilde{U}W^\top)_{\tilde{p}\tilde{q}}\right| = \|\tilde{U}_{\tilde{p}*}\|_2\|\tilde{V}_{\tilde{q}*}\|_2 = \frac{r\sqrt{\mu_0(\tilde{U}\tilde{U}^\top)\mu_0(\tilde{V}\tilde{V}^\top)}}{\sqrt{n_1n_2}}$$

  to obtain $M_1(T_2) = \mu_0(\tilde{U}\tilde{U}^\top)\mu_0(\tilde{V}\tilde{V}^\top)r$.

  Finally, to bound $\left\|\mathcal{P}_{T_2^\perp}(X)\right\|_*$ with $X = U^r\Sigma^r V^{r\top}$ gives

$$\begin{aligned}&\left\|\mathcal{P}_{T_2^\perp}(X)\right\|_* = \|\tilde{U}\tilde{U}^\top X\tilde{V}^\perp\tilde{V}^{\perp\top} + \tilde{U}^\perp\tilde{U}^{\perp\top}X\|_* \\ &\le \sqrt{r}\|\tilde{U}\tilde{U}^\top X\tilde{V}^\perp\tilde{V}^{\perp\top}\|_F + \sqrt{r}\|\tilde{U}^\perp\tilde{U}^{\perp\top}X\|_F \le \sqrt{r}(\sin(\theta_v) + \sin(\theta_u))\|X\|_F.\end{aligned}$$

- Theorem 8, program (4, T_3): notice that any $X \in T_3 \cap S$ can be written as $X = \tilde{U}W^\top$ where the columns of W are arbitrary. Then

$$\left|(\tilde{U}W^\top)_{pq}\right| \le \|\tilde{U}_{p*}\|_2\|W_{q*}\|_2 \le \sqrt{\frac{\mu_0(\tilde{U}\tilde{U}^\top)r}{n_1}},$$

and therefore, $M_1 \leq \mu_0(\tilde{U}\tilde{U}^\top)n_2$. For a lower bound, let $\tilde{p} \in [n_1]$ obtain the maximum row norm of $\tilde{U}$ and choose each row $W_{q*} = c\tilde{U}_{\tilde{p}*}$ were c is a proper normalizing constant achieving $\|W\|_F = 1$. Then

$$M_1 \geq \frac{n_1 n_2}{r}\left|(\tilde{U}W^\top)_{\tilde{p}q}\right|^2 = \frac{n_2\|\tilde{U}_{\tilde{p}*}\|_2^2}{r} = \frac{\mu_0(\tilde{U}\tilde{U}^\top)n_2}{n_1}.$$

$\left\|\mathcal{P}_{T_3^\perp}(U^r\Sigma^r V^{r\top})\right\|_*$ can be bounded as

$$\begin{aligned}&\left\|\mathcal{P}_{T_3^\perp}(U^r\Sigma^r V^{r\top})\right\|_* = \|\tilde{U}^\perp\tilde{U}^{\perp\top}U^r\Sigma^r V^{r\top}\|_* \leq \sqrt{r}\|\tilde{U}^\perp\tilde{U}^{\perp\top}U^r\Sigma^r V^{r\top}\|_F \\ &\leq \sqrt{r}\sin(\theta_u)\|U^r\Sigma^r V^{r\top}\|_F.\end{aligned}$$

- Theorem 10, program $(4, T_4)$: any $X \in T_4 \cap S_{op}$ can be written as $X = X_1 + X_2 + X_3$ where $X_1 \in T_2$, $X_2 = \tilde{U}W^\top$ with range$(W) \subset$ range$(\tilde{V}^\perp)$, and $X_3 = Z\tilde{V}^\top$ with range$(Z) \subset$ range$(\tilde{U}^\perp)$. As before, the largest entry of any X_1 is bounded by $\mu_0 r/\sqrt{n_1 n_2}$. For matrices of the form X_3, write $Z = \tilde{U}^\perp\alpha$ where $\alpha \in \mathbb{C}^{n_1 - r\times r}$ has orthogonal columns and $\|\alpha\|_F \leq 1$. Then,

$$\left|(Z\tilde{V}^\top)_{pq}\right| \leq \|Z_{p*}\|_2\|\tilde{V}_{q*}\|_2 \leq \|\alpha\|\|\tilde{U}^\perp_{p*}\|_2\sqrt{\frac{\mu_0 r}{n_2}} \leq \sqrt{\frac{\mu_0 r(n_1-\mu_2 r)}{n_1 n_2}}.$$

  The last inequality holds since $1 = \|\tilde{U}_{p*}\|_2 + \|\tilde{U}^\perp_{p*}\|_2^2$ and by definition (18). An analogous argument for X_2 and the triangle inequality gives

$$M_1 \leq \mu_0^2 r + \mu_0(n_1 - \mu_2 r) + \mu_0(n_2 - \mu_2 r).$$

  For the lower bound, consider matrices of the form X_3 as before. Choose α with $\|\alpha\|_F = 1$ in an analogous manner to the proof of Theorem 7. Then $M_1 \geq \mu_0(\tilde{V}\tilde{V}^\top)(n_1 - \mu_2 r) \geq n_1 - \mu_2 r$.

  For $\left\|\mathcal{P}_{T_4^\perp}(U^r\Sigma^r V^{r\top})\right\|_*$, it follows that

$$\begin{aligned}&\left\|\mathcal{P}_{T_4^\perp}(U^r\Sigma^r V^{r\top})\right\|_* = \|\tilde{U}^\perp\tilde{U}^{\perp\top}U^r\Sigma^r V^{r\top}\tilde{V}^\perp\tilde{V}^{\perp\top}\|_* \\ &\leq \sqrt{r}\|\tilde{U}^\perp\tilde{U}^{\perp\top}U^r\Sigma^r V^{r\top}\tilde{V}^\perp\tilde{V}^{\perp\top}\|_F \leq \sqrt{r}\sin(\theta_u)\sin(\theta_v)\|U^r\Sigma^r V^{r\top}\|_F.\end{aligned} \tag{33}$$

- Theorem 6, program $(4, T_1)$: recall that $T_1 = \text{span}\{\tilde{U}_{k*}\tilde{V}_{k*}^\top\}_{k\in[r]}$, so every $X \in T_1 \cap S$ can be written as $X = \tilde{U}\Sigma\tilde{V}^\top$ for some diagonal matrix with $\|\Sigma\|_F \leq 1$.

  For the upper bound, it holds that for any $p \in [n_1]$ and $q \in [n_2]$

$$\left|(\tilde{U}\Sigma\tilde{V}^\top)_{pq}\right| = \left|\sum_k \sigma_k\tilde{U}_{pk}\overline{\tilde{V}}_{qk}\right| \leq \left(\sum_k \sigma_k^2\right)^{1/2}\left(\sum_k |\tilde{U}_{pk}|^2|\tilde{V}_{qk}|^2\right)^{1/2} \leq \sqrt{\frac{\mu_1(\tilde{U}\tilde{V}^\top)r}{n_1 n_2}}.$$

  This gives $M_1(T_1) \leq \mu_1(\tilde{U}\tilde{V}^\top)$.

For the lower bound, notice that the singular values can be freely chosen. Let $(\tilde{p}, \tilde{q}) \in [n_1] \times [n_2]$ be such that

$$\left(\sum_k |\tilde{U}_{\tilde{p}k}|^2 |\tilde{V}_{\tilde{q}k}|^2\right)^{1/2} = \sqrt{\frac{\mu_1(\tilde{U}\tilde{V}^\top) r}{n_1 n_2}}.$$

Then choosing $\sigma_k = c\overline{\tilde{U}}_{\tilde{p}k}\tilde{V}_{\tilde{q}k}$ (possibly complex valued which will still be in T_1) where c is a normalization constant (so $\|\Sigma\|_F = 1$) gives

$$\left|(\tilde{U}\Sigma\tilde{V}^\top)_{\tilde{p}\tilde{q}}\right| = \sqrt{\frac{\mu_1(\tilde{U}\tilde{V}^\top) r}{n_1 n_2}},$$

and therefore $M_1(T_1) \geq \mu_1(\tilde{U}\tilde{V}^\top)$ since it is defined as the maximum.

Finally, to bound $\left\|\mathcal{P}_{T_1^\perp}(U^r \Sigma^r {V^r}^\top)\right\|_*$ with $X = U^r \Sigma^r {V^r}^\top$ notice that

$$\mathcal{P}_{T_1^\perp}(X) = \mathcal{P}_{T_2^\perp}(X) + \sum_{k \neq \ell} \tilde{U}_{*k}\tilde{V}_{*\ell}^\top \langle \tilde{U}_{*k}\tilde{V}_{*\ell}^\top, X\rangle,$$

where T_2 is as in Theorem 7 and the term $\|\mathcal{P}_{T_2^\perp}(X)\|_*$ can be bounded as in the proof therein via a triangle inequality. To bound the remaining term, which is a rank r matrix, we obtain

$$\begin{aligned}
&\left\|\sum_{k \neq \ell} \tilde{U}_{*k}\tilde{V}_{*\ell}^\top \langle \tilde{U}_{*k}\tilde{V}_{*\ell}^\top, X\rangle\right\|_* \leq \sqrt{r}\left\|\sum_{k \neq \ell} \tilde{U}_{*k}\tilde{V}_{*\ell}^\top \langle \tilde{U}_{*k}\tilde{V}_{*\ell}^\top, X\rangle\right\|_F \\
&= \sqrt{r}\left(\sum_{k \neq \ell} |\langle \tilde{U}_{*k}\tilde{V}_{*\ell}^\top, X\rangle|^2\right)^{1/2} = \sqrt{r}\left(\sum_{k \neq \ell} |\langle {U^r}^\top \tilde{U}_{*k}\tilde{V}_{*\ell}^\top V^r, \Sigma^r\rangle|^2\right)^{1/2} \\
&\leq \sqrt{r}\|\Sigma^r\|_F \left(\sum_{k \neq \ell} \left\|{U^r}^\top \tilde{U}_{*k}\tilde{V}_{*\ell}^\top V^r\right\|_F^2\right)^{1/2}.
\end{aligned}$$

- <u>Theorem 2, program (1)</u>: as discussed, choose T as T_1 from Theorem 6 but with $\tilde{U} = U^r$ and $\tilde{V} = V^r$. Then $\left\|\mathcal{P}_{T_1^\perp}(U^r \Sigma^r {V^r}^\top)\right\|_* = 0$ and as in the previous case $M_1 = \mu_1(D)$.

  To upper bound M_0, every $X \in T \cap S_{op}$ can be written as $X = U^r \Sigma {V^r}^\top$ for some diagonal matrix with $\|\Sigma\| \leq 1$. Then $M_0(T) \leq \mu_0(D)$, since

$$\begin{aligned}
\|X\|_{2,\infty} &= \max_{k,\ell}\{\|X_{k*}\|_2, \|X_{*\ell}\|_2\} = \max_{k,\ell}\{\|U^r_{k*}\Sigma {V^r}^\top\|_2, \|U^r \Sigma {V^r_{\ell*}}^\top\|_2\} \\
&\leq \max_{k,\ell}\{\|U^r_{k*}\|_2\|\Sigma {V^r}^\top\|, \|U^r\Sigma\|\|V^r_{\ell*}\|_2\} \leq \max_{k,\ell}\{\|U^r_{k*}\|_2, \|V^r_{\ell*}\|_2\} \leq \sqrt{\frac{\mu_0(D) r}{n_2}},
\end{aligned}$$

  where the second inequality holds since U^r, V^r with orthonormal columns and $\|\Sigma\| \leq 1$ give that $\|U^r\Sigma\|$ and $\|\Sigma {V^r}^\top\|$ are bounded by 1.

Appendix C. Proof of Required Lemmas

This section proves the main lemmas required for the proofs in Appendix A: Lemma 11 and Lemma 12.

C.1 Proof of Lemma 11

Lemma 11 is essentially a generalization of dual certificate guarantees for sparse vector recovery to the low-rank matrix recovery case (see Theorem 4.33 by Foucart and Rauhut 2013), and the proof will be similar.

Proof [Proof of Lemma 11] Denote $W = D^\sharp - D$, the goal is to bound $\|W\|_F$. Since $D^\sharp$ is feasible

$$\|\mathcal{A}(W)\|_2 \leq \|\mathcal{A}(D^\sharp) - \mathcal{A}(D) - d\|_2 + \|d\|_2 \leq 2\eta.$$

This will be applied throughout the proof.

Let $Q \in T^\perp$ be such that $\langle D + W, Q\rangle = \|\mathcal{P}_{T^\perp}(D + W)\|_*$ and $G \in T$ be such that $\langle D, G\rangle = \|\mathcal{P}_T(D)\|_*$. By optimality of $D^\sharp$ and feasibility of D,

$$\begin{aligned}
&\|D\|_* \geq \|D^\sharp\|_* = \|D + W\|_* \geq |\langle D + W, G + Q\rangle| \\
=&\left|\langle D + W, G\rangle + \left\|\mathcal{P}_{T^\perp}(D + W)\right\|_*\right| \\
=&\left|\|\mathcal{P}_T(D)\|_* + \langle \mathcal{P}_T(W), G\rangle + \left\|\mathcal{P}_{T^\perp}(D + W)\right\|_*\right| \\
\geq&\|\mathcal{P}_T(D)\|_* - |\langle \mathcal{P}_T(W), G\rangle| + \left\|\mathcal{P}_{T^\perp}(W)\right\|_* - \left\|\mathcal{P}_{T^\perp}(D)\right\|_*.
\end{aligned}$$

Where the second inequality holds by the variational characterization of the nuclear norm, $\|X\|_* = \sup_{\|Y\|\leq 1}\langle X, Y\rangle$.

Using $\|D\|_* \leq \|\mathcal{P}_T(D)\|_* + \|\mathcal{P}_{T^\perp}(D)\|_*$ and rearranging gives

$$\|\mathcal{P}_{T^\perp}(W)\|_* \leq 2\|\mathcal{P}_{T^\perp}(D)\|_* + |\langle \mathcal{P}_T(W), G\rangle|. \tag{34}$$

Introducing $Y = \mathcal{A}^*\mathcal{A}(Z)$, the last term in (34) can be bounded as

$$\begin{aligned}
&|\langle \mathcal{P}_T(W), G\rangle| \leq |\langle \mathcal{P}_T(W), G - \mathcal{P}_T(Y)\rangle| + |\langle \mathcal{P}_T(W), \mathcal{P}_T(Y)\rangle| \\
\leq&\beta_3\|\mathcal{P}_T(W)\|_F + |\langle \mathcal{P}_T(W), \mathcal{P}_T(Y)\rangle| \\
=&\beta_3\|\mathcal{P}_T(W)\|_F + |\langle W - \mathcal{P}_{T^\perp}(W), Y - \mathcal{P}_{T^\perp}(Y)\rangle| \\
=&\beta_3\|\mathcal{P}_T(W)\|_F + |\langle W, Y\rangle - \langle \mathcal{P}_{T^\perp}(W), \mathcal{P}_{T^\perp}(Y)\rangle| \\
\leq&\beta_3\|\mathcal{P}_T(W)\|_F + |\langle W, Y\rangle| + |\langle \mathcal{P}_{T^\perp}(W), \mathcal{P}_{T^\perp}(Y)\rangle|.
\end{aligned} \tag{35}$$

The second equality applies $\langle W, \mathcal{P}_{T^\perp}(Y)\rangle = \langle \mathcal{P}_{T^\perp}(W), \mathcal{P}_{T^\perp}(Y)\rangle = \langle \mathcal{P}_{T^\perp}(W), Y\rangle$.

The three terms in (35) are bounded next, beginning with $\|\mathcal{P}_T(W)\|_F$. The assumed inequalities (21) give that $\|\mathcal{A}(B)\|_2 \geq \beta_1\|B\|_F$, and $\|\mathcal{A}(H)\|_2 \leq \beta_2\|H\|_F$ for any $B \in T$ and $H \in T^\perp$. Therefore,

$$\begin{aligned}
&\|\mathcal{P}_T(W)\|_F \leq \frac{1}{\beta_1}\|\mathcal{A}(\mathcal{P}_T(W))\|_2 \leq \frac{1}{\beta_1}\|\mathcal{A}(W)\|_2 + \frac{1}{\beta_1}\left\|\mathcal{A}(\mathcal{P}_{T^\perp}(W))\right\|_2 \\
\leq&\frac{2\eta}{\beta_1} + \frac{\beta_2}{\beta_1}\|\mathcal{P}_{T^\perp}(W)\|_F.
\end{aligned} \tag{36}$$

The remaining terms in (35), $|\langle \mathcal{P}_{T^\perp}(W), \mathcal{P}_{T^\perp}(Y)\rangle|$ and $|\langle W, Y\rangle|$, can be bounded by assumptions (22)

$$|\langle \mathcal{P}_{T^\perp}(W), \mathcal{P}_{T^\perp}(Y)\rangle| \le \|\mathcal{P}_{T^\perp}(W)\|_* \|\mathcal{P}_{T^\perp}(Y)\| \le \beta_4 \|\mathcal{P}_{T^\perp}(W)\|_*$$

and

$$|\langle W, Y\rangle| \coloneqq |\langle W, \mathcal{A}^*\mathcal{A}(Z)\rangle| = |\langle \mathcal{A}(W), \mathcal{A}(Z)\rangle| \le \|\mathcal{A}(W)\|_2 \|\mathcal{A}(Z)\|_2 \le 2\eta\beta_5.$$

Using these inequalities to bound $|\langle \mathcal{P}_T(W), G\rangle|$ in (34) gives

$$\begin{aligned}\|\mathcal{P}_{T^\perp}(W)\|_* &\le 2\|\mathcal{P}_{T^\perp}(D)\|_* + \frac{2\eta\beta_3}{\beta_1} + \frac{\beta_2\beta_3}{\beta_1}\|\mathcal{P}_{T^\perp}(W)\|_F + \beta_4\|\mathcal{P}_{T^\perp}(W)\|_* + 2\eta\beta_5 \\ &\le 2\|\mathcal{P}_{T^\perp}(D)\|_* + \left(\frac{\beta_2\beta_3}{\beta_1} + \beta_4\right)\|\mathcal{P}_{T^\perp}(W)\|_* + 2\left(\frac{\beta_3}{\beta_1} + \beta_5\right)\eta.\end{aligned}$$

Since by assumption $\rho \coloneqq \frac{\beta_2\beta_3}{\beta_1} + \beta_4 < 1$, rearrange to obtain

$$\|\mathcal{P}_{T^\perp}(W)\|_* \le \frac{2}{1-\rho}\|\mathcal{P}_{T^\perp}(D)\|_* + \frac{2\left(\frac{\beta_3}{\beta_1} + \beta_5\right)\eta}{1-\rho}.$$

From previous calculations, (36),

$$\|\mathcal{P}_T(W)\|_F \le \frac{2\eta}{\beta_1} + \frac{\beta_2}{\beta_1}\|\mathcal{P}_{T^\perp}(W)\|_F \le \frac{2\eta}{\beta_1} + \frac{\beta_2}{\beta_1}\|\mathcal{P}_{T^\perp}(W)\|_*,$$

so that both of these inequalities give

$$\begin{aligned}\|W\|_F &\le \|\mathcal{P}_T(W)\|_F + \|\mathcal{P}_{T^\perp}(W)\|_F \le \|\mathcal{P}_T(W)\|_F + \|\mathcal{P}_{T^\perp}(W)\|_* \\ &\le \frac{2\eta}{\beta_1} + \left(\frac{\beta_2}{\beta_1} + 1\right)\|\mathcal{P}_{T^\perp}(W)\|_* \le C_1\|\mathcal{P}_{T^\perp}(D)\|_* + 2C_2\eta,\end{aligned}$$

with constants given as

$$C_1 \coloneqq 2\left(\frac{\beta_2}{\beta_1} + 1\right)\left(1 - \frac{\beta_2\beta_3}{\beta_1} - \beta_4\right)^{-1},$$

and

$$C_2 \coloneqq \frac{1}{\beta_1} + \left(\frac{\beta_2}{\beta_1} + 1\right)\left(\frac{\beta_3}{\beta_1} + \beta_5\right)\left(1 - \frac{\beta_2\beta_3}{\beta_1} - \beta_4\right)^{-1}.$$

■

C.2 Proof of Lemma 12

The proof of Lemma 12 requires two lemmas, stated here and proven in Section C.3. Before continuing, some useful notation and observations are established for the sampling operators. Recall that M_1 is the ρ-subspace incoherence parameter of T and that the operator that samples with replacement has been normalized as

$$\tilde{\mathcal{A}} \coloneqq \sqrt{\frac{n_1 n_2}{m}}\mathcal{P}_{\tilde{\Omega}},$$

where m and $\tilde{\Omega}$ (with possible repetitions) are defined as in Section A.2.

The operator $\tilde{\mathcal{A}}$ is as an ensemble of matrices $\{\tilde{\mathcal{A}}^k\}_{k\in[m]} \subset \mathbb{C}^{n_1\times n_2}$. Here the superscripts order the matrices such that the action on $X \in \mathbb{C}^{n_1\times n_2}$ is given entry-wise as

$$\tilde{\mathcal{A}}(X)_k = \langle \tilde{\mathcal{A}}^k, X\rangle,$$

for $k \in [m]$. The scaling ensures that $\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}$ forms an isotropic ensemble, that is, for any $X \in \mathbb{C}^{n_1\times n_2}$

$$\mathbb{E}\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}(X) = \sum_{k=1}^{m} \mathbb{E}\tilde{\mathcal{A}}^k\langle \tilde{\mathcal{A}}^k, X\rangle = \sum_{k=1}^{m}\left(\frac{1}{n_1n_2}\sum_{p=1}^{n_1}\sum_{q=1}^{n_2}\frac{n_1n_2}{m}M^{p,q}X_{pq}\right) = X, \tag{37}$$

where $\{M^{p,q}\}_{(p,q)\in[n_1\times n_2]}$ is the canonical $n_1 \times n_2$ matrix basis. Therefore, each $\tilde{\mathcal{A}}^k$ is a random matrix that achieves $\sqrt{\frac{n_1n_2}{m}}M^{p,q}$ with probability $(n_1n_2)^{-1}$.

Next is a lemma that will be useful to establish Lemma 12, in order to apply a concentration inequality. The proof is postponed until Section C.3 and is straightforward from the subspace incoherence assumptions.

Lemma 14 *Define $\tilde{\mathcal{A}}$ as above. Then for $Z \in T$ and all $k \in [m]$*

$$|\langle \tilde{\mathcal{A}}^k, Z\rangle| \leq \|Z\|_F\sqrt{\frac{M_1\rho}{m}} \tag{38}$$

and

$$\mathbb{E}\sum_{k=1}^{m}|\langle \tilde{\mathcal{A}}^k, Z\rangle|^4 \leq \|Z\|_F^4\frac{M_1\rho}{m},$$

where ρ is defined as in (8).

The next Lemma can be considered a generalization of Lemma 3.6 by Rudelson and Vershynin 2008. The argument is due to Liu 2011, but has been tailored to the current setting with a tighter bound in terms of the logarithm degree. Adopting the notation therein, for a matrix $A \in \mathbb{C}^{n_1\times n_2}$ denote $|A)(A|$ as the operator that maps $X \mapsto A\langle A, X\rangle$.

Lemma 15 *Let $m \leq n_1n_2$ and $\epsilon_1, ..., \epsilon_m$ be i.i.d. Rademacher random variables. Then*

$$\begin{aligned}&\mathbb{E}_\epsilon \sup_{X\in T\cap S}\sum_{k=1}^{m}\epsilon_k\langle|\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k|(X), X\rangle \\ \leq&\frac{\tilde{C}\sqrt{M_1\rho}\log(n_1+n_2)\log^{1/2}(m)}{\sqrt{m}}\left(\sup_{X\in T\cap S}\sum_{k=1}^{m}\langle|\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k|(X), X\rangle\right)^{1/2},\end{aligned}$$

where $\tilde{C} > 0$ is an absolute constant.

See Section C.3 for the proof. The proof of Lemma 12 follows.
Proof [Proof of Lemma 12] Let $\mathcal{T} = T \cap S$ and

$$\mathcal{X} := \sup_{X\in\mathcal{T}}\left|\langle(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}} - \mathcal{I})(X), X\rangle\right| = \sup_{X\in\mathcal{T}}\langle(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}} - \mathcal{I})(X), X\rangle,$$

where the last equality holds since $\tilde{\mathcal{A}}^*\tilde{\mathcal{A}} - \mathcal{I}$ is a Hermitian operator.

The goal is to show

$$\mathcal{X} \leq 2\delta,$$

which is achieved proceeding along the lines of Liu 2011; Rauhut 2008; Rudelson and Vershynin 2008. Adopting the notation from Lemma 15 (and Liu 2011), where for a matrix $A \in \mathbb{C}^{n_1 \times n_2}$ the operator $|A)(A|$ maps $X \mapsto A\langle A, X\rangle$ and write

$$\begin{aligned}\mathcal{X} &:= \sup_{X \in \mathcal{T}} \langle (\tilde{\mathcal{A}}^*\tilde{\mathcal{A}} - \mathcal{I})(X), X\rangle := \|\tilde{\mathcal{A}}^*\tilde{\mathcal{A}} - \mathcal{I}\|_{\mathcal{T}} \\ &= \left\| \sum_{k=1}^{m} \left(|\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| - \frac{1}{m}\mathcal{I} \right) \right\|_{\mathcal{T}} \\ &= \left\| \sum_{k=1}^{m} \left(|\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| - \mathbb{E}|\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| \right) \right\|_{\mathcal{T}},\end{aligned}$$

where the last equality holds due to isotropy of the ensemble (37) with i.i.d. samples. $\mathbb{E}\mathcal{X}$ will be bounded first, and then a concentration inequality will be applied to show this random variable is concentrated around its mean.

Using symmetrization (as in equation (42) of Liu 2011, which uses Lemma 6.3 by Ledoux and Talagrand 1991, gives

$$\mathbb{E}\mathcal{X} \leq 2\mathbb{E}_{\tilde{\Omega}}\mathbb{E}_{\epsilon} \left\| \sum_{k=1}^{m} \epsilon_k |\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| \right\|_{\mathcal{T}},$$

where ϵ_k are Rademacher random variables. Applying Lemma 15, which requires $m \leq n_1 n_2$, gives

$$\mathbb{E}_{\epsilon} \left\| \sum_{k} \epsilon_k |\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| \right\|_{\mathcal{T}} \leq C_1 \left\| \sum_{k} |\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| \right\|_{\mathcal{T}}^{1/2},$$

where

$$C_1 := \frac{\tilde{C}\sqrt{M_1\rho}\log(n_1+n_2)\log^{1/2}(m)}{\sqrt{m}},$$

and $\tilde{C} > 0$ is an absolute constant given in Lemma 15. Summarizing and continuing these calculations,

$$\begin{aligned}\mathbb{E}\mathcal{X} &\leq 2C_1\mathbb{E}\left(\left\| \sum_{k} |\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| \right\|_{\mathcal{T}} \right)^{1/2} \\ &\leq 2C_1\mathbb{E}\left(\left\| \sum_{k} \left(|\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| - \frac{1}{m}\mathcal{I} \right) \right\|_{\mathcal{T}} + 1 \right)^{1/2} \\ &\leq 2C_1\left(\mathbb{E}\left\| \sum_{k} \left(|\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| - \frac{1}{m}\mathcal{I} \right) \right\|_{\mathcal{T}} + 1 \right)^{1/2} \\ &= 2C_1\left(\mathbb{E}\mathcal{X} + 1\right)^{1/2}.\end{aligned}$$

Therefore

$$\frac{\mathbb{E}\mathcal{X}}{\sqrt{\mathbb{E}\mathcal{X}+1}} \leq \frac{2\tilde{C}\sqrt{M_1\rho}\log(n_1+n_2)\log^{1/2}(m)}{\sqrt{m}} \leq \frac{2\sqrt{2}\tilde{C}\sqrt{M_1\rho}\log^{3/2}(n_1+n_2)}{\sqrt{m}},$$

where the last inequality holds since $m \leq (n_1+n_2)^2$ by assumption. Given $\delta > 0$, $\mathbb{E}\mathcal{X} \leq \delta$ if

$$\frac{2\sqrt{2}\tilde{C}\sqrt{M_1\rho}\log^{3/2}(n_1+n_2)}{\sqrt{m}} \leq \frac{\delta}{\sqrt{\delta+1}}. \tag{39}$$

A concentration inequality will show that $\mathcal{X}$ is close to its expected value with high probability.

Theorem 16 (Theorem 8.42 in (Foucart and Rauhut 2013)) *Let $\mathcal{F}$ be a countable set of functions $f : \mathbb{C}^{n_1\times n_2} \mapsto \mathbb{R}$. Let $Y_1, ..., Y_m$ be independent random matrices in $\mathbb{C}^{n_1\times n_2}$ such that $\mathbb{E}f(Y_k) = 0$ and $f(Y_k) \leq K$ almost surely for all $k \in [m]$ and for all $f \in \mathcal{F}$. Define Z as the random variable*

$$Z = \sup_{f\in F}\sum_{k=1}^{m} f(Y_k).$$

Let $\sigma^2 > 0$ be such that $\mathbb{E}\sum_{k=1}^{m} f(Y_\ell)^2 \leq \sigma^2$ for all $f \in \mathcal{F}$. Then for all $\delta \geq 0$

$$\mathbb{P}\left(Z \geq \mathbb{E}Z + \delta\right) \leq \exp\left(-\frac{\delta^2}{2\sigma^2 + 4K\mathbb{E}Z + 2\delta K/3}\right).$$

To apply the theorem, let $X \in \mathcal{T}$ generate a set of functions $f_X : \mathbb{C}^{n_1\times n_2} \to \mathbb{R}$ via

$$f_X(Z) \coloneqq |\langle Z, X\rangle|^2 - \frac{1}{m}.$$

Then notice that

$$\begin{aligned}\mathcal{X} &\coloneqq \left\|\sum_k \left(|\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| - \frac{1}{m}\mathcal{I}\right)\right\|_{\mathcal{T}} \\ &\coloneqq \sup_{X\in\mathcal{T}}\sum_k \left\langle \left(|\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k| - \frac{1}{m}\mathcal{I}\right)(X), X\right\rangle \\ &= \sup_{X\in\mathcal{T}}\sum_k f_X\left(\tilde{\mathcal{A}}^k\right) = \sup_{X\in\tilde{\mathcal{T}}}\sum_k f_X\left(\tilde{\mathcal{A}}^k\right),\end{aligned}$$

where $\tilde{\mathcal{T}}$ is a dense countable subset of $\mathcal{T}$.

For all $k \in [m]$ and $X \in \mathcal{T}$, by the first part of Lemma 14

$$f_X\left(\tilde{\mathcal{A}}^k\right) \leq |\langle\tilde{\mathcal{A}}^k, X\rangle|^2 \leq \frac{M_1\rho}{m},$$

and $K = \frac{M_1\rho}{m}$ from Theorem 16.

Now for σ^2, apply the second part of Lemma 14 and isometry of the ensemble to obtain

$$\mathbb{E}\sum_k f_X\left(\tilde{\mathcal{A}}^k\right)^2 \leq \mathbb{E}\sum_k |\langle \tilde{\mathcal{A}}^k, X\rangle|^4 \leq \frac{M_1\rho}{m} := \sigma^2.$$

To finish, assuming

$$\sqrt{m} \geq 2\sqrt{2}\tilde{C}\sqrt{M_1\rho}\log^{3/2}(n_1+n_2)\frac{\sqrt{1+\delta}}{\delta}, \tag{40}$$

gives $\mathbb{E}\mathcal{X} \leq \delta$ according to (39) and by Theorem 16

$$\mathcal{X} \leq \mathbb{E}\mathcal{X} + \delta < 2\delta$$

with probability of failure not exceeding

$$\exp\left(-\frac{m\delta^2}{2M_1\rho + 4M_1\rho\delta + 2M_1\rho\delta/3}\right) \leq \exp\left(-\frac{6m\delta^2}{19M_1\rho}\right).$$

The last inequality holds assuming $\delta \leq \frac{1}{4}$, under which (40) holds if

$$\sqrt{m} \geq \frac{\sqrt{10}\tilde{C}\sqrt{M_1\rho}\log^{3/2}(n_1+n_2)}{\delta},$$

where the statement of the theorem absorbs all the absolute constants into C.

■

C.3 Proof of Additional Lemmas

This section supplies the proofs of Lemmas 13, 14 and 15.

Proof [Proof of Lemma 13] The main ingredient is a matrix Bernstein inequality (Theorem 1.6 by Tropp 2012). As in Section C.2, expand

$$\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}(Z) - Z = \sum_{k=1}^m \left(\tilde{\mathcal{A}}^k\langle\tilde{\mathcal{A}}^k, Z\rangle - \frac{1}{m}Z\right)$$

which is a sum of independent and centered random matrices. In order to apply the Bernstein inequality, the operator norms of each summand and the matrix variance statistic need to be bounded.

Notice that for any $k \in [m]$

$$\left\|\tilde{\mathcal{A}}^k\langle\tilde{\mathcal{A}}^k, Z\rangle - \frac{1}{m}Z\right\| \leq \frac{n_1n_2}{m}\|Z\|_\infty := R$$

and

$$\mathbb{E}\sum_k \left(\tilde{\mathcal{A}}^k\langle\tilde{\mathcal{A}}^k, Z\rangle - \frac{1}{m}Z\right)\left(\tilde{\mathcal{A}}^k\langle\tilde{\mathcal{A}}^k, Z\rangle - \frac{1}{m}Z\right)^\top = \sum_k \left(\mathbb{E}\tilde{\mathcal{A}}^k\tilde{\mathcal{A}}^{k\top}|\langle\tilde{\mathcal{A}}^k, Z\rangle|^2 - \frac{1}{m^2}ZZ^\top\right)$$
$$= \frac{n_1n_2}{m}M_1 - \frac{1}{m}ZZ^\top,$$

where $M_1 \in \mathbb{C}^{n_1 \times n_1}$ is a diagonal matrix whose entries are the diagonal elements of $ZZ^\top$. Therefore,

$$\begin{aligned}
&\left\| \mathbb{E} \sum_k \left(\tilde{\mathcal{A}}^k \langle \tilde{\mathcal{A}}^k, Z \rangle - \frac{1}{m} Z \right) \left(\tilde{\mathcal{A}}^k \langle \tilde{\mathcal{A}}^k, Z \rangle - \frac{1}{m} Z \right)^\top \right\| \leq \frac{1}{m} \left(n_1 n_2 \|M_1\| + \|ZZ^\top\| \right) \\
&= \frac{1}{m} \left(n_1 n_2 \max_{1 \leq k \leq n_1} \|Z_{k*}\|_2^2 + \|ZZ^\top\| \right) \leq \frac{2 n_1 n_2}{m} \|Z\|_{\infty,2}^2 \coloneqq \sigma^2.
\end{aligned}$$

The last inequality holds by Gershgorin circle theorem, which gives that for some $k \in [n_1]$

$$\|ZZ^\top\| \leq |(ZZ^\top)_{kk}| + \sum_{\ell \neq k} |(ZZ^\top)_{k\ell}| = \|Z_{k*}\|_2^2 + \sum_{\ell \neq k} |\langle Z_{k*}, Z_{\ell*} \rangle| \leq n_1 \|Z\|_{\infty,2}^2.$$

Analogously, bound

$$\left\| \mathbb{E} \sum_k \left(\tilde{\mathcal{A}}^k \langle \tilde{\mathcal{A}}^k, Z \rangle - \frac{1}{m} Z \right)^\top \left(\tilde{\mathcal{A}}^k \langle \tilde{\mathcal{A}}^k, Z \rangle - \frac{1}{m} Z \right) \right\| \leq \sigma^2.$$

Apply Theorem 1.6 in by Tropp 2012 with R, σ^2 above and

$$t = \frac{4 n_1 n_2 \log(n_1 + n_2) \|Z\|_\infty / 3}{2m} + \frac{\sqrt{16 n_1^2 n_2^2 \log^2(n_1 + n_2) \|Z\|_\infty^2 / 9 + 32 m n_1 n_2 \log(n_1 + n_2) \|Z\|_{\infty,2}^2}}{2m}$$

to obtain the desired probability of success. The statement of the lemma simplifies the upper bound on the operator norm by noting that

$$t \leq \frac{4 n_1 n_2 \log(n_1 + n_2) \|Z\|_\infty / 3 + 2\sqrt{2 m n_1 n_2 \log(n_1 + n_2)} \|Z\|_{\infty,2}}{m}.$$

■

Lemma 14, which admits a straightforward proof.

Proof [Proof of Lemma 14] The inequality (38) is straightforward by definition and scaling. For the remaining claim, if $Z \in T$ then by (38)

$$\begin{aligned}
&\mathbb{E} \sum_{k=1}^m |\langle \tilde{\mathcal{A}}^k, Z \rangle|^4 = \frac{n_1 n_2}{m} \sum_{p=1}^{n_1} \sum_{q=1}^{n_2} |Z_{pq}|^4 \\
&\leq \frac{n_1 n_2}{m} \left(\max_{(p,q) \in [n_1] \times [n_2]} |Z_{pq}|^2 \right) \sum_{p=1}^{n_1} \sum_{q=1}^{n_2} |Z_{pq}|^2 \leq \|Z\|_F^4 \frac{M_1 \rho}{m}.
\end{aligned}$$

■

The proof of Lemma 15 is due to Liu 2011, tailored here to fit the specific setting. This modified argument results in a tighter bound in terms of the logarithmic dependency.

Adopting the author's notation, in what follows for a matrix $A \in \mathbb{C}^{n_1 \times n_1}$ denote $|A)(A|$ as the operator that maps $X \mapsto A\langle A, X\rangle$.

Proof [Proof of Lemma 15] The argument will rely on the work of Liu 2011 for brevity, referring the reader to the proof of Lemma 3.1 in Section A therein. With U_2 defined by Liu 2011, notice that $T \cap S \subset U_2$ since every matrix in T is rank ρ (where ρ is defined in (8)). The result here is obtained in a similar manner, but considering non-square matrices and linear subspace T as Banach space in its own right to compute its covering number. To this end, it is important notice that T equipped with the Frobenius norm is a Banach space and $\{\tilde{\mathcal{A}}^k\}_{k=1}^m \subset T^*$, where T^* denotes the dual space of T, have dual norm bounded as

$$\|\tilde{\mathcal{A}}^k\|_{T^*} = \sup_{X \in T \cap S} \left|\langle \tilde{\mathcal{A}}^k, X\rangle\right| \leq \sqrt{\frac{M_1 \rho}{m}} := \sqrt{\rho} K, \qquad \forall k \in [m] \tag{41}$$

by Lemma 14. Notice that $K := \sqrt{M_1/m}$.

With this in mind, proceed as in the proof of Lemma 3.1 by Liu 2011 (with $T \cap S$ replacing U_2) up to equation (15) which is obtained via comparison principle to a Gaussian process and Dudley's inequality. Combined with bound (18) therein and a change of variables shows

$$\mathbb{E}_\epsilon \sup_{X \in T \cap S} \sum_{k=1}^m \epsilon_k \langle |\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k|(X), X\rangle \leq 48\sqrt{2\pi} R \sqrt{\rho} \int_0^\infty \log^{1/2} \mathcal{N}\left(\frac{1}{\sqrt{\rho}}\left(T \cap S\right), \|\cdot\|_X, \epsilon\right) d\epsilon,$$

where $\mathcal{N}\left(B, \|\cdot\|, \epsilon\right)$ is the number of balls of radius ϵ in a metric $\|\cdot\|$ needed to cover a set B,

$$R := \left(\sup_{X \in T \cap S} \sum_{k=1}^m \langle |\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k|(X), X\rangle\right)^{1/2}$$

and $\|\cdot\|_X$ is a semi-norm defined for $Y \in \mathbb{C}^{n_1 \times n_2}$ as

$$\|Y\|_X = \max_{k \in [m]} |\langle \tilde{\mathcal{A}}^k, Y\rangle|.$$

To bound the integral, bound $\mathcal{N}\left(\frac{1}{\sqrt{\rho}}\left(T \cap S\right), \|\cdot\|_X, \epsilon\right)$ in two different ways. For $Y \in \frac{1}{\sqrt{\rho}}\left(T \cap S\right)$, notice that $\|Y\|_X \leq K$ by (41) so that

$$\mathcal{N}\left(\frac{1}{\sqrt{\rho}}\left(T \cap S\right), \|\cdot\|_X, \epsilon\right) \leq \mathcal{N}\left(K \cdot B_X, \|\cdot\|_X, \epsilon\right), \tag{42}$$

where B_X is the unit ball in $\|\cdot\|_X$. For small ϵ and with $n_1 \geq n_2$, use (42) and equation (20) by Liu 2011 to obtain

$$\mathcal{N}\left(\frac{1}{\sqrt{\rho}}\left(T \cap S\right), \|\cdot\|_X, \epsilon\right) \leq \left(1 + \frac{2K}{\epsilon}\right)^{2n_1^2}. \tag{43}$$

For large ϵ, apply Lemma 3.2 by Liu 2011 (Lemma 1 by Guédon et al. 2008) with $E = T$ equipped with the Frobenious norm to obtain

$$\mathcal{N}\left(\frac{1}{\sqrt{\rho}}\left(T \cap S\right), \|\cdot\|_X, \epsilon\right) = \mathcal{N}\left(T \cap S, \|\cdot\|_X, \epsilon\sqrt{\rho}\right) \leq \exp\left(\frac{C_1^2 K^2 \log(m)}{\epsilon^2}\right), \tag{44}$$

where C_1 is an absolute constant given by Maurey's empirical method. The inequality holds by (41) and since T has modulus of convexity of power type 2 with constant $\lambda(T) = 1$ and dual space type 2 constant $T_2(T^*) \leq 1$ due to the Frobenius norm (see Theorem A3 and A4 in the Appendix of Aubrun 2009).

To bound the integral, split it at $A \coloneqq K/n_1$ and use (43) for small ϵ to obtain

$$\begin{aligned}
&\int_0^A \log^{1/2} \mathcal{N}\left(\frac{1}{\sqrt{\rho}}\left(T \cap S\right), \|\cdot\|_X, \epsilon\right) d\epsilon \leq \int_0^A \sqrt{2} n_1 \log^{1/2}\left(1 + \frac{2K}{\epsilon}\right) d\epsilon \\
&\leq \sqrt{2} n_1 \int_0^A \left(1 + \log\left(1 + \frac{2K}{\epsilon}\right)\right) d\epsilon = \sqrt{2} n_1 \int_{1/A}^{\infty} \left(1 + \log\left(1 + 2Ky\right)\right) \frac{dy}{y^2} \\
&\leq \sqrt{2} n_1 \int_{1/A}^{\infty} \left(1 + \log\left((A + 2K)y\right)\right) \frac{dy}{y^2} \leq \sqrt{2} K (2 + \log(1 + 2n_1)),
\end{aligned}$$

where the final bound holds by integrating $\log\left((A + 2K)y\right)/y^2$ by parts.

Consider the remaining part of the integral up to K, since when $\epsilon > K$ by (42) it holds that $\mathcal{N}\left(\frac{1}{\sqrt{\rho}}\left(T \cap S\right), \|\cdot\|_X, \epsilon\right) = 1$. Using (44) gives

$$\int_A^K \log^{1/2} \mathcal{N}\left(\frac{1}{\sqrt{\rho}}\left(T \cap S\right), \|\cdot\|_X, \epsilon\right) d\epsilon \leq \int_A^K \frac{C_1 K \log^{1/2}(m)}{\epsilon} d\epsilon = C_1 K \log^{1/2}(m) \log(n_1).$$

In conclusion

$$\begin{aligned}
&\mathbb{E}_\epsilon \sup_{X \in T \cap S} \sum_{k=1}^m \epsilon_k \langle |\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k|(X), X\rangle \\
&\leq 48\sqrt{2\pi} R \sqrt{\rho} \left(\sqrt{2} K (2 + \log(1 + 2n_1)) + C_1 K \log^{1/2}(m) \log(n_1)\right) \\
&\leq \frac{\tilde{C}\sqrt{M_1 \rho} \log^{1/2}(m) \log(n_1 + n_2)}{\sqrt{m}} \left(\sup_{X \in T \cap S} \sum_{k=1}^m \langle |\tilde{\mathcal{A}}^k)(\tilde{\mathcal{A}}^k|(X), X\rangle\right)^{1/2},
\end{aligned}$$

for some absolute constant $\tilde{C}$.

■

The main difference in this proof as opposed to the proof of Lemma 3.1 by Liu 2011 is that the containment of $\frac{1}{\sqrt{\rho}}\left(T \cap S\right)$ in the nuclear norm ball B_1 is not considered here. Rather, the linear subspace T is viewed as a Banach space with unit ball $T \cap S$. This allows for a direct application of Lemma A.3, reducing the logarithmic dependency by a factor of $\log^{3/2}(n_1)$. The same gain does not seem straightforward for the universal result of Liu 2011. Otherwise, using the arguments here, the sample complexity of Pauli measurements can be reduced to $\mathcal{O}(n_1 \rho \log^{3/2}(n_1))$ when universality is not imposed (for example, via a dual certificate instead of the restricted isometry property).

References

Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization.

Journal of Machine Learning Research, 10(29):803–826, 2009. URL `http://jmlr.org/papers/v10/abernethy09a.html`.

Aleksandr Aravkin, Rajiv Kumar, Hassan Mansour, Ben Recht, and Felix J. Herrmann. Fast methods for denoising matrix completion formulations, with applications to robust seismic data interpolation. *SIAM Journal on Scientific Computing*, 36(5):S237–S266, 2014. doi: 10.1137/130919210. URL `https://doi.org/10.1137/130919210`.

Guillaume Aubrun. On almost randomizing channels with a short kraus decomposition. *Communications in Mathematical Physics*, 288(3):1103–1116, 2009. doi: 10.1007/s00220-008-0695-y.

Siavash Bayat and Sajad Daei. An optimal hybrid nuclear norm regularization for matrix sensing with subspace prior information. *IEEE Access*, 8:130937–130946, 2020. doi: 10.1109/ACCESS.2020.3009688.

Emmanuel Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):1615–3383, apr 2009. ISSN 0001-0782. doi: https://doi.org/10.1007/s10208-009-9045-5. URL `https://doi.org/10.1145/2184319.2184343`.

Emmanuel J. Candes and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010. doi: 10.1109/TIT.2010.2044061.

Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015. doi: 10.1109/TIT.2015.2415195.

Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Coherent matrix completion. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page I–674–I–682. JMLR.org, 2014.

Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Completing any low-rank matrix, provably. *Journal of Machine Learning Research*, 16(94):2999–3034, 2015. URL `http://jmlr.org/papers/v16/chen15b.html`.

Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. Matrix completion with noisy side information. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper/2015/file/0609154fa35b3194026346c9cac2a248-Paper.pdf`.

Zlatko Drmac. On principal angles between subspaces of euclidean space. *SIAM Journal on Matrix Analysis and Applications*, 22(1):173–194, 2000. doi: 10.1137/S0895479897320824. URL `https://doi.org/10.1137/S0895479897320824`.

Armin Eftekhari, Michael B Wakin, and Rachel A Ward. MC2: a two-phase algorithm for leveraged matrix completion. *Information and Inference: A Journal of the IMA*, 7(3):581–604, 02 2018a. ISSN 2049-8764. doi: 10.1093/imaiai/iax020. URL `https://doi.org/10.1093/imaiai/iax020`.

Armin Eftekhari, Dehui Yang, and Michael B. Wakin. Weighted matrix completion and recovery with prior subspace information. *IEEE Transactions on Information Theory*, 64 (6):4044–4071, 2018b. doi: 10.1109/TIT.2018.2816685.

M. Fazel, H. Hindi, and S.P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148)*, volume 6, pages 4734–4739 vol.6, 2001. doi: 10.1109/ACC.2001.945730.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing.* Birkhäuser Basel, 2013. ISBN 0817649476.

Rina Foygel and Nathan Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 315–340, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL `https://proceedings.mlr.press/v19/foygel11a.html`.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011. doi: 10.1109/TIT.2011.2104999.

Olivier Guédon, Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Majorizing measures and proportional subsets of bounded orthonormal systems. *Revista Matemática Iberoamericana*, 24(3):1075 – 1095, 2008. doi: rmi/1228834305. URL `https://doi.org/`.

Prateek Jain and Inderjit S. Dhillon. Provable inductive matrix completion. *CoRR*, abs/1306.0626, 2013. URL `http://arxiv.org/abs/1306.0626`.

Sun Ji-guang. Perturbation of angles between linear subspaces. *Journal of Computational Mathematics*, 5(1):58–61, 1987. ISSN 02549409, 19917139. URL `http://www.jstor.org/stable/43692289`.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer-Verlag, 1991.

Yi-Kai Liu. Universal low-rank matrix recovery from pauli measurements. *NIPS'11 Proceedings of the 24th International Conference on Neural Information Processing Systems*, abs/1103.2816:1638–1646, 2011.

Yong Luo, Tongliang Liu, Dacheng Tao, and Chao Xu. Multiview matrix completion for multilabel image classification. *IEEE Transactions on Image Processing*, 24(8):2355–2368, 2015. doi: 10.1109/TIP.2015.2421309.

Oscar López, Rajiv Kumar, Özgür Yılmaz, and Felix J. Herrmann. Off-the-grid low-rank matrix recovery and seismic data reconstruction. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):658–671, 2016. doi: 10.1109/JSTSP.2016.2555482.

Duc Minh Nguyen, Evaggelia Tsiligianni, and Nikos Deligiannis. Extendable neural matrix completion. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6328–6332, 2018. doi: 10.1109/ICASSP.2018.8462164.

Holger Rauhut. Stability results for random sampling of sparse trigonometric polynomials. *IEEE Transactions on Information Theory*, 54(12):5661–5670, 2008. doi: 10.1109/TIT.2008.2006382.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(104):3413–3430, 2011. URL `http://jmlr.org/papers/v12/recht11a.html`.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. doi: 10.1137/070697835. URL `https://doi.org/10.1137/070697835`.

Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008. doi: https://doi.org/10.1002/cpa.20227. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20227`.

Nathan Srebro and Tommi S. Jaakkola. *Learning with Matrix Factorizations*. PhD thesis, USA, 2004. AAI0807530.

Nathan Srebro and Russ R Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL `https://proceedings.neurips.cc/paper/2010/file/67d96d458abdef21792e6d8e590244e7-Paper.pdf`.

Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, 9(2):137–154, 1992.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays . *Bioinformatics*, 17(6):520–525, 06 2001. ISSN 1367-4803. doi: 10.1093/bioinformatics/17.6.520. URL `https://doi.org/10.1093/bioinformatics/17.6.520`.

Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019. doi: 10.1137/18M1183480. URL `https://doi.org/10.1137/18M1183480`.

Ewout van den Berg and Michael P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2009. doi: 10.1137/080714488. URL `https://doi.org/10.1137/080714488`.

Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL `https://proceedings.neurips.cc/paper/2013/file/e58cc5ca94270acaceed13bc82dfedf7-Paper.pdf`.

Yi Yang, Jianwei Ma, and Stanley Osher. Seismic data reconstruction via matrix completion. *Inverse Problems and Imaging*, 7(4):1379–1392, 2013.

Jinfeng Yi, Tianbao Yang, Rong Jin, Anil K. Jain, and Mehrdad Mahdavi. Robust ensemble clustering by matrix completion. In *2012 IEEE 12th International Conference on Data Mining*, pages 1176–1181, 2012. doi: 10.1109/ICDM.2012.123.

Jinfeng Yi, Lijun Zhang, Rong Jin, Qi Qian, and Anil K. Jain. Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1400–III–1408. JMLR.org, 2013.

Yijun Zhang, Shashin Sharan, and Felix J. Herrmann. *High-frequency wavefield recovery with weighted matrix factorizations*, pages 3959–3963. 2019. doi: 10.1190/segam2019-3215103.1. URL `https://library.seg.org/doi/abs/10.1190/segam2019-3215103.1`.

Yijun Zhang, Shashin Sharan, Oscar López, and Felix J. Herrmann. *Wavefield recovery with limited-subspace weighted matrix factorizations*, pages 2858–2862. 2020. doi: 10.1190/segam2020-3428365.1. URL `https://library.seg.org/doi/abs/10.1190/segam2020-3428365.1`.

Bo Zhao, Justin P. Haldar, Cornelius Brinegar, and Zhi-Pei Liang. Low rank matrix recovery for real-time cardiac mri. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 996–999, 2010. doi: 10.1109/ISBI.2010.5490156.