Symmetric Matrix Completion with ReLU Sampling

Huikang Liu [*1] **Peng Wang** [*2] **Longxiu Huang** [3] **Qing Qu** [2] **Laura Balzano** [2]

Abstract

We study the problem of symmetric positive semi-definite low-rank matrix completion (MC) with deterministic entry-dependent sampling. In particular, we consider rectified linear unit (ReLU) sampling, where only positive entries are observed, as well as a generalization to threshold-based sampling. We first empirically demonstrate that the landscape of this MC problem is not globally benign: Gradient descent (GD) with random initialization will generally converge to stationary points that are not globally optimal. Nevertheless, we prove that when the matrix factor with a small rank satisfies mild assumptions, the nonconvex objective function is geodesically strongly convex on the quotient manifold in a neighborhood of a planted low-rank matrix. Moreover, we show that our assumptions are satisfied by a matrix factor with i.i.d. Gaussian entries. Finally, we develop a tailor-designed initialization for GD to solve our studied formulation, which empirically always achieves convergence to the global minima. We also conduct extensive experiments and compare MC methods, investigating convergence and completion performance with respect to initialization, noise level, dimension, and rank.

1. Introduction

Low-rank matrix completion (MC) refers to the task of filling in the missing entries of a partially observed low-rank matrix. It has found applications in diverse fields, such as recommendation systems (Koren et al., 2009), sequential bioinformatics (Zheng et al., 2013), and computer vision (Ji et al., 2010), to name a few. In particular, symmetric positive semi-definite (PSD) low-rank MC has applications in covariance matrix completion (Candes & Plan, 2010; Hosseini & Sebt, 2017), Hankel matrix completion (Chen & Chi, 2013; Usevich & Comon, 2016; Cai et al., 2023), and Euclidean distance matrix completion (Laurent, 2001; Al-Homidan & Wolkowicz, 2005; Dokmanic et al., 2015). An extensive body of literature investigates the statistical and optimization properties of the low-rank MC problem using different approaches, such as nuclear-norm minimization (Recht, 2011; Candes & Plan, 2010; Candes & Recht, 2012), spectral method (Keshavan et al., 2010; Chatterjee, 2015), and matrix factorization (Sun & Luo, 2016), among others. To analyze the MC problem, most of these approaches rely on the following assumptions: (1) the underlying matrix is low-rank and incoherent, and (2) the entries are observed *independent of the matrix entries* according to a probabilistic mechanism, e.g., uniformly at random. Based on these assumptions, it is possible to prove how many observed entries are required so that the missing entries can be exactly or approximately completed.

The majority of existing work assumes that the entries are observed uniformly at random, independent of the underlying matrix values. However, this assumption is strict and often violated in practice. In real-world applications where data are collected from measurements, such as distance matrices, missing entries tend to be those that are harder to collect. When data are collected from participants, such as online shopping or surveys, missing answers are typically highly correlated with the question and the value of the true answer. For example, surveys about sensitive topics will have missing entries on any culturally or morally problematic answers. Moreover, online ratings tend to be skewed to the high end; for example, most people do not read a book or watch a movie that they might dislike. Even in sensor systems, missing data are not likely to be independent and completely random. Sensors may saturate at a certain value or break down based on environmental conditions that also affect the sensor value. In all these applications, the probability of missing entries in a matrix is dependent on the underlying values, sometimes deterministically.

Despite being highly relevant for applications, the problem of recovering missing entries when the sampling mechanism depends on the entries remains a challenging and relatively under-explored area of research. Existing papers along these

[*]Equal contribution [1]Antai College of Economics and Management, Shanghai Jiao Tong University [2]Department of Electrical Engineering and Computer Science, University of Michigan [3]Department of Computational Mathematics, Science and Engineering & Department of Mathematics, Michigan State University. Correspondence to: Laura Balzano <girasole@umich.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

lines give impractical results, such as those where the recovery metric provides no guarantees for recovering entries that are never observed (Foygel et al., 2011; Foucart et al., 2020), or high-dimensional consistency results that do not admit finite sample guarantees or make strong assumptions on the sampling probability functions, for example that they are Lipschitz in the matrix entry value and nonzero everywhere (Bhattacharya & Chatterjee, 2022). Research works on MC with deterministic sampling focused on understanding fundamental properties that the sampling patterns must exhibit (Lee et al., 2023). However, none of these works provide clear and practical guidelines for what MC problems can be solved in their given settings and for what algorithms will successfully complete them.

1.1. Our Contributions

Our work aims to advance the understanding of MC with deterministic sampling by focusing on symmetric PSD low-rank MC with ReLU sampling, where only the positive entries of a matrix are observed. In this setting, under relatively mild assumptions, we first show that the globally optimal solutions of the low-rank MC problem with a known rank can exactly (resp. approximately) complete the missing entries in the noiseless (resp. noisy) setting. Moreover, we prove that the objective function is geodesically strongly convex on the quotient manifold around the *planted* low-rank matrix, i.e., the underlying true low-rank matrix. Therefore, with an initial point close enough to the planted low-rank matrix, GD will converge to this desired matrix. This motivates us to tailor an initialization method for GD. We empirically demonstrate that GD with our initialization always converges to the planted low-rank matrix. It is worth mentioning that this objective function landscape result is the first of its kind within the literature on dependent or deterministic sampling for MC.

While ReLU sampling is a specific setting, it has the potential to be generalized to a broad class of common missing data problems, e.g., where only a range of values is observed. In this direction, we have also provided more general assumptions for a broader class of sampling functions. For example, a threshold sample where we observe all entries above a positive threshold will also lead to a strongly convex objective around the planted low-rank matrix. This more general setting requires stronger assumptions, but it gives us more general results that also apply to the noisy setting. We note that this thresholding sampling setting generally means that a constant fraction of entries are observed. This is an interesting sampling regime when observations are entry-dependent, and a useful one for many sensor data applications where it may be feasible to collect a moderate number of matrix entries.

Even in the best case of our theorems, empirical results (shown in Section 4) drastically outperform the theory, leaving an exciting open question as to whether better assumptions and proof techniques can deliver theory that matches empirical observation.

1.2. Literature Review

Due to space constraints, we defer detailed related work to Appendix B and focus here on the most related works. Bishop & Yu (2014) prove PSD matrix recovery for their proposed algorithm with deterministic sampling of principal submatrices. Our assumptions are different and our results showing the benign local landscape of the problem are applicable to any gradient-based algorithm. Bhattacharya & Chatterjee (2022) assumes that each entry of a low-rank matrix $\boldsymbol{M}^*$ is observed with some probability $f(\boldsymbol{M}^*)$, where f is applied entry-wise and is essentially Lipschitz and nonzero. Their proofs apply to the high-dimensional regime where the dimensions grow but rank is uniformly bounded, as in (Chatterjee, 2020).

In an empirical work that studies ReLU sampling among others (Naik et al., 2022), the authors conclude that convex methods generally do not work as well as nonconvex in the dependent-sampling setting, motivating our focus on the nonconvex objective function. Finally, the work in (Saul, 2022; Seraghiti et al., 2023) both seek a non-negative low-rank matrix $\boldsymbol{M}^*$ such that a sparse $\boldsymbol{M} = f(\boldsymbol{M}^*)$ and f is a given nonlinearity, such as ReLU. They provide multiple algorithms for learning $\boldsymbol{M}^*$ (such as EM, alternating block descent) but no theoretical guarantees for when this low-rank approximation exists.

Notation. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we use $\sigma_{\max}(\boldsymbol{A})$ or $\|\boldsymbol{A}\|$ to denote its largest singular value (i.e., spectral norm), $\sigma_{\min}(\boldsymbol{A})$ its smallest non-zero singular value, $\|\boldsymbol{A}\|_F$ its Frobenius norm, and a_{ij} its (i,j)-th element. Given a vector $\boldsymbol{a} \in \mathbb{R}^n$, we denote its Euclidean norm by $\|\boldsymbol{a}\|$ and the i-th entry by a_i. Given a positive integer n, we denote by $[n]$ the set $\{1,\dots,n\}$. Let $\mathcal{O}^{m\times n} = \left\{\boldsymbol{Q} \in \mathbb{R}^{m\times n} : \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_n\right\}$ denote the set of all $m \times n$ orthonormal matrices. In particular, let $\mathcal{O}^m = \{\boldsymbol{Q} \in \mathbb{R}^{m\times m} : \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_m\}$ denote the set of all $m \times m$ orthogonal matrices. Given an integer n, we define $[n] := \{1,\dots,n\}$.

2. Problem Formulation

MC with ReLU sampling. In this work, we consider a noisy symmetric PSD MC completion problem. Specifically, let $\boldsymbol{M}^\star := \boldsymbol{U}^\star\boldsymbol{U}^{\star T} \in \mathbb{R}^{n\times n}$ be a symmetric PSD matrix, where $\boldsymbol{U}^\star \in \mathbb{R}^{n\times r}$. In addition, let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a noisy version of $\boldsymbol{M}^\star$ generated by

$$\boldsymbol{M} := \boldsymbol{M}^\star + \boldsymbol{\Delta}, \tag{1}$$

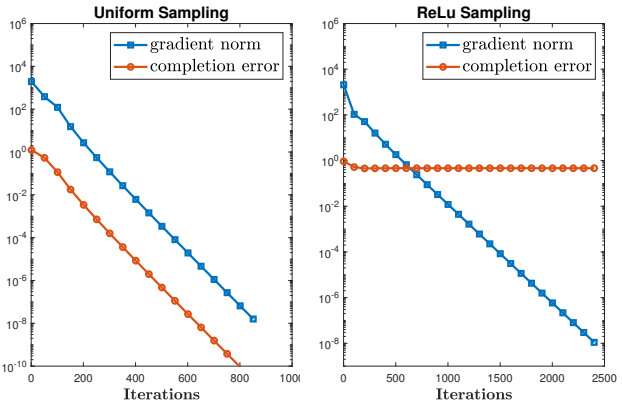


Figure 1. **Recovery and convergence performance of GD for solving the MC problem with the uniform** $(p = 0.2)$ **and ReLU sampling in the noiseless case.** We apply GD with Gaussian random initialization for solving Problem (3) with the uniform and ReLU sampling, respectively. Then, we plot the gradient norm (i.e., $\|\nabla F(\boldsymbol{U}^{(t)})\|_F$) and completion error (i.e., $\|\boldsymbol{U}^{(t)}\boldsymbol{U}^{(t)^T} - \boldsymbol{M}\|_F/\|\boldsymbol{M}\|_F$) against number of iterations.

where $\boldsymbol{\Delta} \in \mathbb{R}^{n\times n}$ is a noise matrix. It is worth noting that $\boldsymbol{M}$ is non-symmetric when the noise matrix $\boldsymbol{\Delta}$ is not symmetric.

There are many applications where it is common to observe only a partial set of the entries of $\boldsymbol{M}$ (Nguyen et al., 2019). This could be due to data collection, experimental constraints, or inherent missing information (Hu et al., 2008). In this work, we consider a setting where the missingness pattern of the matrix is dependent on the underlying values in the matrix and is deterministic given the matrix entries. Specifically, we suppose that only the non-negative entries in $\boldsymbol{M}$ can be observed, i.e., the observed set

$$\Omega = \{(i,j) \in [n] \times [n] : m_{ij} \geq 0\}. \tag{2}$$

Notably, this sampling regime is commonly referred to as ReLU sampling in the literature (see, e.g., Naik et al. (2022); Mazumdar & Rawat (2018)), as it utilizes the function $f(x) = \max\{0, x\}$, known as the rectified linear unit (ReLU) function in deep learning[1]. Then, our goal is to complete the missing entries of $\boldsymbol{M}^\star$ from the observed entries in $\boldsymbol{M}_\Omega$.

Before proceeding, we make some remarks on this MC problem. First, existing works (Naik et al., 2022; Saul, 2022; Seraghiti et al., 2023) have empirically studied the low-rank MC problem with ReLU sampling and focused on proposing efficient algorithms to address this problem. In particular, Mazumdar & Rawat (2018) has theoretically studied the recovery performance of the ReLU-based representation learning problem under a probabilistic model. This differs from our work, which studies the global optimality and optimization landscape of the MC problem. Second, this sampling is merely a specific instance within a broader set of general deterministic sampling schemes. We prove results on one such generalization, where we observe values above or below a threshold. We believe our work will be a springboard for the study of more general practical entry-dependent sampling schemes.

Optimization formulation. Leveraging the low-rank structure in Equation (1), we consider the following non-convex MC problem to complete the missing entries of $\boldsymbol{M}^*$:

$$\min_{\boldsymbol{U}\in\mathbb{R}^{n\times r}} F(\boldsymbol{U}) := \frac{1}{4}\left\|(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{M})_\Omega\right\|_F^2. \tag{3}$$

Note that the rank of the optimization variable $\boldsymbol{U}$ is exactly the rank of the planted low-rank matrix $\boldsymbol{M}^*$. In the noiseless setting with uniform sampling and incoherence assumptions, Ge et al. (2016) have shown that Problem (3) has a benign global optimization landscape in the sense that it has no spurious local minima—all local minima must also be global minima. This, together with the result in (Lee et al., 2016), implies that gradient descent (GD) with random initialization with high probability converges to globally optimal solutions that achieve exact completion; see Figure 1(a). One may then conjecture that Problem (3) in the ReLU sampling setting also has such a benign optimization landscape. However, this conjecture is refuted by empirical evidence in Figure 1(b). This result is typical in the ReLU sampling setting, illustrating that GD with random initialization may converge to a spurious critical point that is not globally optimal. The next question is whether it is even possible to recover the missing entries by applying GD for this setting. In this work, we answer this question in the affirmative.

3. Main Results

In this section, we first present our theoretical result on symmetric PSD low-rank MC with ReLU sampling in the noiseless case under mild assumptions in Section 3.1. In Section 3.2, we extend these results to the noisy case under more general sampling assumptions. In Section 3.3, we show that all the introduced assumptions hold with high probability when the entries of $\boldsymbol{U}^\star$ are i.i.d. sampled from the standard Gaussian distribution.

3.1. Noiseless Case and ReLU Sampling

In this subsection, we consider the noiseless case, i.e., $\boldsymbol{\Delta} = \boldsymbol{0}$, and ReLU sampling in (2). We start by introducing some assumptions on the underlying matrix $\boldsymbol{U}^\star$ and the sampling set Ω. Noting that the r-dimensional space has 2^r orthants, we denote by $\mathcal{C}_i$ the i-th orthant of the r-dimensional space

[1] See (Nair & Hinton, 2010) for early use of the phrase "rectified linear unit," but it had been in use for neural nets long before, referred to by the name "linear threshold unit" (Wersing et al., 2001) or "positive part" (Jarrett et al., 2009), among others.

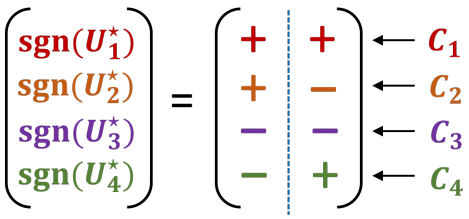


Figure 2. **An illustrative figure on the partition of rows of $\boldsymbol{U}^\star \in \mathbb{R}^{n\times 2}$.** We rearrange the rows of $\boldsymbol{U}^\star$ and partition them into 4 blocks, each belonging to different orthants.

for each $i \in [2^r]$. For example, there are 4 orthants (i.e., quadrants) in the 2-dimensional space and 8 orthants (i.e., octants) in the 3-dimensional space. Here, $\mathcal{C}_i$ for each $i \in [2^r]$ is ordered such that the signs of components of a vector belonging to orthant $\mathcal{C}_i$ differ from those in orthant $\mathcal{C}_{i+1}$ in only one component. For ease of exposition, without loss of generality, we assume the rows of $\boldsymbol{U}^\star \in \mathbb{R}^{n\times r}$ are partitioned into the following blocks

$$\boldsymbol{U}^\star = \begin{bmatrix} \boldsymbol{U}_1^{\star T} & \boldsymbol{U}_2^{\star T} & \dots & \boldsymbol{U}_{2^r}^{\star T} \end{bmatrix}^T \in \mathbb{R}^{n\times r}, \quad (4)$$

where each row of $\boldsymbol{U}_i^\star \in \mathbb{R}^{n_i\times r}$ belongs to the i-th orthant $\mathcal{C}_i$ for each $i \in [2^r]$ and $\sum_{i=1}^{2^r} n_i = n$. For example, when $r = 2$, the rows of different blocks of $\boldsymbol{U}_i^*$ takes the signs as shown in Figure 2. Based on the above setup, we make the following assumption.

Assumption 3.1. For each $i \in [2^r]$, $\mathrm{rank}(\boldsymbol{U}_i^\star) = r$.

We remark that if r is much smaller than $\log n$ and the entries of $\boldsymbol{U}^\star$ are i.i.d. sampled from a distribution symmetric about zero, such as the standard normal distribution, then the generated submatrices are full rank with high probability.

Next, we discuss the second assumption that will lead to unique completion. This assumption is illustrated in Figure 3. For any pair of indices $(i, j) \in [2^r] \times [2^r]$, we denote by $\boldsymbol{M}^{(i,j)} \in \mathbb{R}^{n_i\times n_j}$ the (i, j)-th block of $\boldsymbol{M}$, i.e., $\boldsymbol{M}^{(i,j)} = \boldsymbol{U}_i^\star \boldsymbol{U}_j^{\star T}$. Moreover, we denote the set of observations in the (i, j) block by $\Omega_{i,j} = \{(k, l) \in [n_i] \times [n_j] : m_{kl}^{(i,j)} \geq 0\}$ where $m_{kl}^{(i,j)}$ is the (k, l)-th entry of $\boldsymbol{M}^{(i,j)}$. Let a column vector $\boldsymbol{u}_{i,k}^\star \in \mathbb{R}^r$ denote the k-th row of $\boldsymbol{U}_i^\star$ for each $k \in [n_i]$. Then, we have $m_{kl}^{(i,j)} = \boldsymbol{u}_{i,k}^{\star T} \boldsymbol{u}_{j,l}^\star$. It is obvious that $\Omega_{i,i} = [n_i] \times [n_i]$ under the ReLU sampling since $\boldsymbol{u}_{i,k}^{\star T} \boldsymbol{u}_{i,l}^\star \geq 0$ always holds for all $k, l \in [n_i]$ due to the fact that the rows of $\boldsymbol{U}_i^\star$ have the same sign. Besides, since $\boldsymbol{M}^{(i+1,i)} = \boldsymbol{U}_{i+1}^\star \boldsymbol{U}_i^{\star T}$ and the signs of components in each row of $\boldsymbol{U}_{i+1}^\star$ and those of $\boldsymbol{U}_i^\star$ only differ in one component, one can image that there are enough observations in $\Omega_{i+1,i}$ with ReLU sampling. More precisely, we can formalize the above observation as follows:

Assumption 3.2. For any $i \in [2^r - 1]$, we have $|\Omega_{i+1,i}| \geq r^2$ and the matrix space spanned by $\{\boldsymbol{u}_{i+1,k}^\star \boldsymbol{u}_{i,l}^{\star T} : (k, l) \in \Omega_{i+1,i}\}$ is the whole space $\mathbb{R}^{r\times r}$.

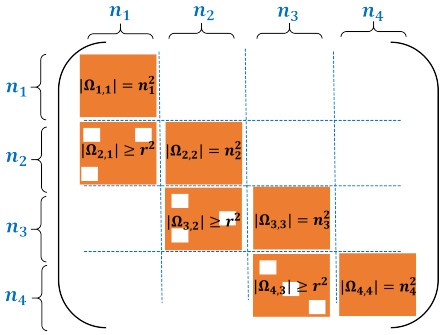


Figure 3. **An illustrative figure on Assumption 3.2 when $r = 2$.** Orange pixels denote observed entries, while white pixels denote missing entries.

Assumptions 3.1 and 3.2 guarantee that there are enough positive (observed) entries for the uniqueness of the completion. We note that these assumptions can be checked from the observed matrix by finding a permutation of rows and columns such that the diagonal blocks are fully observed and of rank r and the off-diagonal blocks have enough entries. In Section 3.3, we will show that, when $\boldsymbol{U}^\star$ is i.i.d. Gaussian random matrix and $r \leq \frac{1}{2}\log n$, Assumptions 3.1 and 3.2 hold with high probability.

Characterization of global optimality. Based on the two assumptions, we are ready to characterize the global optimality of the MC problem (3) under ReLU sampling.

Theorem 3.3. *Suppose that $\boldsymbol{\Delta} = \boldsymbol{0}$ in (1), the observed set Ω is defined in (2), and Assumptions 3.1 and 3.2 hold. Then, $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ is a global optimal solution to Problem (3) if and only if it satisfies $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{M}^\star$.*

We defer the proof of this theorem to Appendix A.1.1. This theorem demonstrates that even under the ReLU sampling regime, the global optimal solutions of Problem (3) still recover the underlying matrix $\boldsymbol{M}^\star$ exactly, just as in the uniform sampling regime.

Uniqueness of global solutions on the manifold. Remark that the objective function $F(\cdot)$ of Probelm (3) is invariant under any orthogonal matrices in $\mathbb{R}^{r\times r}$, i.e., $F(\boldsymbol{U}\boldsymbol{Q}) = F(\boldsymbol{U})$ for all $\boldsymbol{Q} \in \mathcal{O}^r$. To characterize this invariance to orthogonal transformations, let $\sim$ denote an equivalent relation on $\mathbb{R}^{n\times r}$ with the equivalence class

$$[\boldsymbol{U}] := \left\{\boldsymbol{V} \in \mathbb{R}^{n\times r} : \boldsymbol{V} \sim \boldsymbol{U} \text{ iff } \exists \boldsymbol{Q} \in \mathcal{O}^r, \boldsymbol{V} = \boldsymbol{U}\boldsymbol{Q}\right\}.$$

Then, we define the Riemannian quotient manifold $\mathcal{M}$ as

$$\mathcal{M} := \mathbb{R}^{n\times r}\backslash \sim = \{[\boldsymbol{U}] : \boldsymbol{U} \in \mathbb{R}^{n\times r}\}. \quad (5)$$

If we consider Problem (3) on the manifold $\mathcal{M}$, Theorem 3.3 implies that Problem (3) has a unique global optimal solution $[\boldsymbol{U}^\star]$. Next, we will show that the objective function $F(\cdot)$ exhibits geodesic strong convexity on the quotient manifold $\mathcal{M}$ near the global optimum $[\boldsymbol{U}^\star]$.

Main technical ingredient. The proof of Theorem 3.3 relies on the following technical lemma, which could be independent interest. This lemma indicates that the distance between two low-rank matrices on the quotient manifold $\mathcal{M}$ is bounded by their subspace distance. Its proof is also deferred to Appendix A.1.3.

Lemma 3.4. *For arbitrary $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times r}$ with $r \leq n$, suppose that rank$(\boldsymbol{U}) = r$. Then, there exists an orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}^r$ such that*

$$\|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{Q}\|_F \leq \frac{\sigma_{\min}(\boldsymbol{U}) + \|\boldsymbol{V}\|}{\sigma^2_{\min}(\boldsymbol{U})}\|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T\|_F.$$

Preliminary setup for manifold optimization. As discussed above, we study Problem (3) over the quotient manifold $\mathcal{M}$ defined in (5). Toward this end, we introduce some concepts of manifold optimization, such as tangent space, Riemannian gradient, and Riemannian Hessian. Since the formal definition of the *tangent space* to a quotient manifold is abstract, we describe it informally as follows. We define the *vertical space* at $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ denoted by $\mathcal{S}_{\boldsymbol{U}}$ (see Boumal (2023, Definition 9.23, Example 9.25)) as

$$\mathcal{S}_{\boldsymbol{U}} := \left\{\boldsymbol{U}\boldsymbol{R} \in \mathbb{R}^{n\times r} : \boldsymbol{R} + \boldsymbol{R}^T = \boldsymbol{0}, \boldsymbol{R} \in \mathbb{R}^{r\times r}\right\}. \quad (6)$$

As shown in (14), the orthogonal complement of $\mathcal{S}_{\boldsymbol{U}}$, which is called the horizontal space and denoted by $\mathcal{S}^{\perp}_{\boldsymbol{U}}$, is

$$\mathcal{S}^{\perp}_{\boldsymbol{U}} := \left\{\boldsymbol{D} \in \mathbb{R}^{n\times r} : \boldsymbol{U}^T\boldsymbol{D} = \boldsymbol{D}^T\boldsymbol{U}\right\}. \quad (7)$$

According to Boumal (2023, Definition 9.24)), for any $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, there exists a bijective mapping $\mathrm{lift}_{\boldsymbol{U}}$ between any tangent vector $\boldsymbol{\xi} \in \mathrm{T}_{[\boldsymbol{U}]}\mathcal{M}$ and any matrix $\boldsymbol{D} \in \mathcal{S}^{\perp}_{\boldsymbol{U}}$, i.e.,

$$\mathrm{lift}_{\boldsymbol{U}} : \mathrm{T}_{[\boldsymbol{U}]}\mathcal{M} \mapsto \mathcal{S}^{\perp}_{\boldsymbol{U}},\ \mathrm{lift}_{\boldsymbol{U}}(\boldsymbol{\xi}) = \boldsymbol{D},$$

where $\mathrm{T}_{[\boldsymbol{U}]}\mathcal{M} : \mathcal{M} \to \mathbb{R}^{n\times r}$ denotes the tangent space to $\mathcal{M}$ at $[\boldsymbol{U}]$. According to Boumal (2023, Propositions 9.38 and 9.44)), the Riemannian gradient and Hessian at $[\boldsymbol{U}]$ along a direction $\boldsymbol{\xi} \in \mathrm{T}_{[\boldsymbol{U}]}\mathcal{M}$, denoted by $\mathrm{grad}F([\boldsymbol{U}])[\boldsymbol{\xi}]$ and $\mathrm{Hess}F([\boldsymbol{U}])[\boldsymbol{\xi},\boldsymbol{\xi}]$, are computed by

$$\begin{aligned}\mathrm{grad}F([\boldsymbol{U}])[\boldsymbol{\xi}] &= \langle \nabla F(\boldsymbol{U}), \mathrm{lift}_{\boldsymbol{U}}(\boldsymbol{\xi})\rangle, \\ \mathrm{Hess}F([\boldsymbol{U}])[\boldsymbol{\xi},\boldsymbol{\xi}] &= \left\langle \nabla^2 F(\boldsymbol{U})[\mathrm{lift}_{\boldsymbol{U}}(\boldsymbol{\xi})], \mathrm{lift}_{\boldsymbol{U}}(\boldsymbol{\xi})\right\rangle. \quad (8)\end{aligned}$$

Geodesic strong convexity on $\mathcal{M}$. Equipped with the above setup, we analyze the optimization landscape of Problem (3) around the global optimal solutions in the ReLU sampling regime. Although Problem (3) does not possess a benign *global* optimization landscape, we show that it has a favorable *local* optimization landscape.

Theorem 3.5. *Under Assumptions 3.1 and 3.2, $F(\cdot)$ is geodesically strongly convex on the quotient manifold $\mathcal{M}$ at $[\boldsymbol{U}^\star]$, i.e., for all $\boldsymbol{\xi} \in \mathrm{T}_{[\boldsymbol{U}^\star]}\mathcal{M}$,*

$$\mathrm{Hess}F([\boldsymbol{U}^\star])[\boldsymbol{\xi},\boldsymbol{\xi}] \geq \frac{\gamma}{2}\|\mathrm{lift}_{\boldsymbol{U}^\star}(\boldsymbol{\xi})\|^2_F, \quad (9)$$

where

$$\gamma := \min_{\boldsymbol{D}\in\mathcal{S}^{\perp}_{\boldsymbol{U}^\star}, \|\boldsymbol{D}\|_F=1} \|(\boldsymbol{U}^\star\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{U}^{\star T})_\Omega\|^2_F > 0. \quad (10)$$

We defer the proof of this theorem to Appendix A.1.2. Intuitively, this theorem demonstrates that the objective function $F(\cdot)$ is strongly convex on the manifold $\mathcal{M}$ in the neighborhood of $\boldsymbol{U}^\star$. Consequently, if a tailored initialization in the local neighborhood of $\boldsymbol{U}^\star$ is available, GD is guaranteed to find a global optimal solution.

From geodesic strong convexity on $\mathcal{M}$ to strong convexity in Euclidean space. Notably, the above geodesic strong convexity on the manifold $\mathcal{M}$ implies strong convexity of $F(\cdot)$ along some directions in Euclidean space. Specifically, for any $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, let $\boldsymbol{U}^T\boldsymbol{U}^\star = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^T$ be a singular value decomposition of $\boldsymbol{U}^T\boldsymbol{U}^\star$, where $\boldsymbol{P}, \boldsymbol{Q} \in \mathcal{O}^r$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, and $\widetilde{\boldsymbol{U}} := \boldsymbol{U}\boldsymbol{P}\boldsymbol{Q}^T$. Using this and (9), we can show

$$\nabla^2 F(\boldsymbol{U}^\star)[\widetilde{\boldsymbol{U}} - \boldsymbol{U}^\star, \widetilde{\boldsymbol{U}} - \boldsymbol{U}^\star] \geq \frac{\gamma}{2}\|\widetilde{\boldsymbol{U}} - \boldsymbol{U}^\star\|^2_F.$$

Please refer to Appendix A.1 for detailed proof.

3.2. Noisy Case and General Deterministic Sampling

In this subsection, we extend our above results to the noisy case with a general deterministic sampling regime. Toward this goal, we need the following assumption on the sampling pattern, which generalizes Assumptions 3.1 and 3.2.

Assumption 3.6. There exists a collection of index set $\{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_K\}$ such that the following conditions hold:

(a) $\bigcup_{k\in[K]} \mathcal{I}_k = [n]$;

(b) $\mathcal{I}_k \times \mathcal{I}_k \subseteq \Omega$ holds for all $k \in [K]$;

(c) There exists $\lambda > 0$ such that $\boldsymbol{U}^{\star T}_k \boldsymbol{U}^\star_k \succeq \lambda \boldsymbol{I}_r$ for each $k \in [K]$, where $\boldsymbol{U}^\star_k$ is the matrix with the rows consisting of $\boldsymbol{u}^\star_i$ for all $i \in \mathcal{I}_k$;

(d) For any $k \neq l \in [K]$, there exists a path $k_0 \to k_1 \to \cdots \to k_s$ such that $k_0 = k$, $k_s = l$ and $\mathcal{I}_{k_{j-1}} \times \mathcal{I}_{k_j} \subseteq \Omega$ for all $j \in [s]$.

Now, let us explain these conditions in more detail. Condition (a) guarantees that each row of $\boldsymbol{U}^\star$ belongs to at least one submatrix; Condition (b) guarantees that the diagonal blocks of $\boldsymbol{M}$ index by $\mathcal{I}_k$ for all $k \in [K]$ are fully observed; Condition (c) indicates that there are enough rows in each part such that the submatrix $\boldsymbol{U}^\star_k$ is of full row rank; Finally, condition (d) ensures that there are some off-diagonal blocks that are also fully observed and any two of these blocks are connected via a path. We note that the last condition is strict but aids significantly in the proof of Theorem 3.8. With a

slightly more careful analysis, this condition could be weakened so that off-diagonal blocks are only partially observed. In Section 3.3, we show that under the setting where the entries of $\boldsymbol{U}^\star$ are i.i.d. sampled from the standard Gaussian distribution, $r \le O(\log n)$, and the noise is bounded, Assumption 3.6 holds with high probability.

Comparison to (Bishop & Yu, 2014). Assumption 3.6 is similar to the assumptions in (Bishop & Yu, 2014), but as far as we know, neither implies the other. More specifically, Bishop & Yu (2014) consider deterministic sampling and assume that a collection of subsets of Ω admit an ordering such that any two adjacent parts have enough overlap. However, in the ReLU sampling setting, this order is hard to construct, and it is not clear whether there exists such an ordered collection of subsets. Moreover, both the analysis and the algorithm proposed in (Bishop & Yu, 2014) heavily rely on these ordered subsets, while we only use our partition for analysis.

General sampling regimes. Notably, Assumption 3.6 can be applied to more general sampling regimes. For example, consider a positive-threshold sampling regime, where the entry m_{ij} of $\boldsymbol{M}$ is observed if $m_{ij} \ge \eta$ for a constant $\eta \ge 0$. In particular, ReLU sampling corresponds to the case $\eta = 0$.

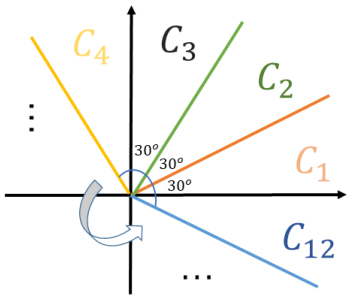


Figure 4. **A figure on the partition of the plane into 12 equal sectors.**

Example 3.7 (Positive-Threshold Mask)**.** Consider the noiseless case with $r = 2$ and the entry m_{ij} is observed if $m_{ij} \ge \eta$ for a threshold $\eta > 0$. We partition the 2D plane into 12 equal sectors, shown in Figure 4. Let us define a collection of index sets as follows:

$$\mathcal{I}_k := \{i \in [n] : \boldsymbol{u}_i^\star \in \mathcal{C}_k\},\ \forall k \in [12].$$

Obviously, we have $[n] = \cup_{k\in[K]}\mathcal{I}_k$, and thus condition (a) in Assumption 3.6 holds. For any $i, j \in \mathcal{I}_k$, the rows $\boldsymbol{u}_i^\star$ and $\boldsymbol{u}_j^\star$ belong to the same sector $\mathcal{C}_k$, which implies that the angle between $\boldsymbol{u}_i^\star$ and $\boldsymbol{u}_j^\star$ is less than $\pi/6$. Therefore, we have $m_{ij} = \boldsymbol{u}_i^{\star T}\boldsymbol{u}_j^\star = \|\boldsymbol{u}_i^\star\|\|\boldsymbol{u}_j^\star\|\cos(\pi/6) = \sqrt{3}\|\boldsymbol{u}_i^\star\|\|\boldsymbol{u}_j^\star\|/2$. As long as the threshold $\eta \le \min\{\sqrt{3}\|\boldsymbol{u}_i^\star\|\|\boldsymbol{u}_j^\star\|/2 : i, j \in \mathcal{C}_k,\ \forall k \in [12]\}$, the entry m_{ij} can be observed for all $i, j \in \mathcal{I}_k$, and thus $\mathcal{I}_k \times \mathcal{I}_k \subseteq \Omega$. This implies that condition (b) in Assumption 3.6 holds. Condition (c) in Assumption 3.6 holds as long as each $\mathcal{I}_k$ contains at least two elements $i \ne j$ such that $\boldsymbol{u}_i^\star$ is not parallel to $\boldsymbol{u}_j^\star$. Finally, for any $i \in \mathcal{I}_k$ and $j \in \mathcal{I}_{k+1}$, one can verify that $m_{ij} \ge \|\boldsymbol{u}_i^\star\|\|\boldsymbol{u}_j^\star\|/2$ using the similar argument. Consequently, as long as the threshold $\eta \le \min\{\|\boldsymbol{u}_i^\star\|\|\boldsymbol{u}_j^\star\|/2 : i \in \mathcal{I}_k,\ j \in \mathcal{I}_{k+1},\ \forall k \in [11]\}$, we have $\mathcal{I}_k \times \mathcal{I}_{k+1} \subseteq \Omega$. That is, all the near-diagonal blocks of M will be fully observed, and thus condition (d) in Assumption 3.6 holds.

Global optimality and local landscape analysis. Based on Assumption 3.6, we are ready to characterize the global optimality and local optimization landscape of Problem (3) in the noisy case.

Theorem 3.8. *Let $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ be any global optimal solution to Problem (3). Under Assumption 3.6, we have*

$$\|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}^\star\boldsymbol{U}^{\star T}\|_F \le \frac{c}{\lambda}\|\boldsymbol{\Delta}\|_F$$

where $c > 0$ depends on λ, $\|\boldsymbol{U}^\star\|_F$ and $\|\boldsymbol{\Delta}\|_F$.

We defer the proof of the theorem to Appendix A.2.1. In particular, the dependence of c on λ, $\|\boldsymbol{U}^\star\|_F$ and $\|\boldsymbol{\Delta}\|_F$ in specified in (39). Notably, this theorem generalizes Theorem 3.3 to the noisy case with general deterministic sampling satisfying Assumption 3.6.

Theorem 3.9. *Let $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ be any global optimal solution to Problem (3). Suppose that the Assumption 3.6 holds and the noise matrix $\boldsymbol{\Delta}$ in (1) satisfies*

$$\|\boldsymbol{\Delta}\|_F \le \min\left\{\frac{\gamma}{8}, \frac{\lambda\sqrt{\lambda}}{4(\sqrt{\lambda}+\sqrt{\gamma}+\|\boldsymbol{U}^\star\|)}, \frac{\lambda\sqrt{\gamma}}{4\tilde{c}_0\left(1+\kappa(\sqrt{\gamma}+2\|\boldsymbol{U}^\star\|)\right)}\right\} \tag{11}$$

where γ is provided in (10) in Theorem 3.5, $\tilde{c}_0 > 0$ depends on $\gamma, \boldsymbol{U}^\star$, and λ, and $\kappa > 0$ only depends on $\boldsymbol{U}^\star$. Then, $F(\cdot)$ is geodesically strongly convex on $\mathcal{M}$ at $[\boldsymbol{U}]$, i.e., for all $\boldsymbol{\xi} \in \mathrm{T}_{[\boldsymbol{U}]}\mathcal{M}$,

$$\mathrm{Hess}F([\boldsymbol{U}])[\boldsymbol{\xi},\boldsymbol{\xi}] \ge \left(\frac{\gamma}{8} - \|\boldsymbol{\Delta}\|_F\right)\|\mathrm{lift}_{\boldsymbol{U}}(\boldsymbol{\xi})\|_F^2.$$

We defer the proof of this theorem to Appendix A.2.2. Intuitively, this theorem demonstrates that the objective function $F(\cdot)$ is strongly convex on the quotient manifold $\mathcal{M}$ in the neighborhood of $[\boldsymbol{U}]$.

3.3. Verification of Assumptions

In this subsection, we show that when the entries of $\boldsymbol{U}^\star$ are i.i.d. sampled from the standard Gaussian distribution, Assumptions 3.1, 3.2, and 3.6 all hold with high probability using a concentration argument.

Noiseless setting. We first show that when $\boldsymbol{\Delta} = \mathbf{0}$ in (1), if the entries of $\boldsymbol{U}^\star$ are i.i.d. Gaussian random variables and the rank is small, Assumptions 3.1 and 3.2 hold with a high probability.

Proposition 3.10. *Suppose that $r \leq (\log_2 n)/2$ and the entries of $\boldsymbol{U}^\star$ are i.i.d. sampled from the standard Gaussian distribution. Then, Assumptions 3.1 and 3.2 hold with probability at least $1 - \sqrt{n}\exp(-\sqrt{n}/16)$.*

We defer the proof of this proposition to Appendix A.3.1. While the rank assumption is strict, we empirically observe in Section 4.4 that it can be relaxed. This rank requirement arises from our proof technique. It is of great interest to find an alternative approach to improve our analysis technique.

Noisy setting. For the noisy case, i.e., $\boldsymbol{\Delta} \neq \mathbf{0}$ in (1), we construct the collection of index sets mentioned in Assumption 3.6 in a similar manner to the partition shown in Figure 4. However, it is more difficult to characterize it in high dimensions. To address this, we introduce the concept of ϵ-net (see Definition A.2) to construct these index sets. Under the Gaussian distribution and ReLU sampling, we prove the following proposition that shows as long as $r \leq O(\log n)$ and the noise is small enough, Assumption 3.6 hold with high probability. We defer the proof to Appendix A.3.2.

Proposition 3.11. *Suppose that $r \leq \log n/(4\log 3)$, the entries of $\boldsymbol{U}^\star$ are i.i.d. sampled from the standard Gaussian distribution, and the noise matrix $\boldsymbol{\Delta}$ satisfies $\|\boldsymbol{\Delta}\|_\infty < \min\left\{\|\boldsymbol{u}_k^\star\|^2 : k \in [n]\right\}/2$. Then, Assumption 3.6 holds with probability at least $1 - 1/n$.*

4. Experimental Results

In this section, we conduct numerical experiments to validate our theoretical developments and demonstrate the performance of several state-of-the-art algorithms on the MC problem with ReLU sampling. All of our experiments are implemented in MATLAB R2023a on a PC with 32GM memory and Intel(R) Core(TM) i7-11800H 2.3GHz CPU. Experiments with Euclidean distance matrix completion can be found in Appendix C.3.

4.1. Our Proposed Algorithm

A natural approach to solving Problem (3) is to apply GD as follows: Given the current iterate $\boldsymbol{U}^{(t)}$, we generate the next iterate via

$$\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} - \eta\nabla F(\boldsymbol{U}^{(t)}),\ t \geq 0, \tag{12}$$

where $\boldsymbol{Z} = (\boldsymbol{U}\boldsymbol{U}^T)_\Omega - \boldsymbol{M}_\Omega$ and $\nabla F(\boldsymbol{U}) = \left(\boldsymbol{Z} + \boldsymbol{Z}^T\right)\boldsymbol{U}$. According to Theorem 3.5, Problem (3) with the ReLU sampling exhibits a benign *local* optimization landscape, even though it possesses a complicated *global* landscape as illustrated in Figure 1. Consequently, GD may not be effective for solving Problem (3) unless a proper initial point $\boldsymbol{U}^{(0)}$ is available. This motivates us to propose a tailor-designed initialization as follows: We first randomly generate a matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times r}$, whose entries are i.i.d. sampled from the standard normal distribution. Then, we construct $\boldsymbol{Q} = \boldsymbol{Y}\boldsymbol{Y}^T$ and generate a new matrix $\bar{\boldsymbol{Q}} \in \mathbb{R}^{n\times n}$ via

$$\bar{q}_{ij} = \begin{cases} m_{ij}, & \text{if } (i,j) \in \Omega, \\ -|q_{ij}|, & \text{otherwise.} \end{cases} \tag{13}$$

Finally, we apply a truncated SVD to $\bar{\boldsymbol{Q}}$ to obtain an initial point $\boldsymbol{U}^{(0)} \in \mathcal{O}^{n\times r}$, where the columns of $\boldsymbol{U}^{(0)}$ consist of right singular vectors of $\bar{\boldsymbol{Q}}$ corresponding to its top r singular values. Notably, we leverage the low-rank structure of $\boldsymbol{M}$ and the ReLU sampling to design the initialization scheme. According to this and (12), we now summarize our proposed method for solving Problem (3) in Algorithm 1.

Algorithm 1 GD for MC with ReLU sampling

Input: observed set Ω, observed matrix $\boldsymbol{M}_\Omega$
Generate $\bar{\boldsymbol{Q}} \in \mathbb{R}^{n\times n}$ using (13)
Apply a truncated SVD to $\bar{\boldsymbol{Q}}$ to obtain $\boldsymbol{U}^{(0)} \in \mathcal{O}^{n\times r}$
for $t = 0, 1, 2, \ldots$ **do**
  $\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} - \eta\nabla F(\boldsymbol{U}^{(t)})$
end for

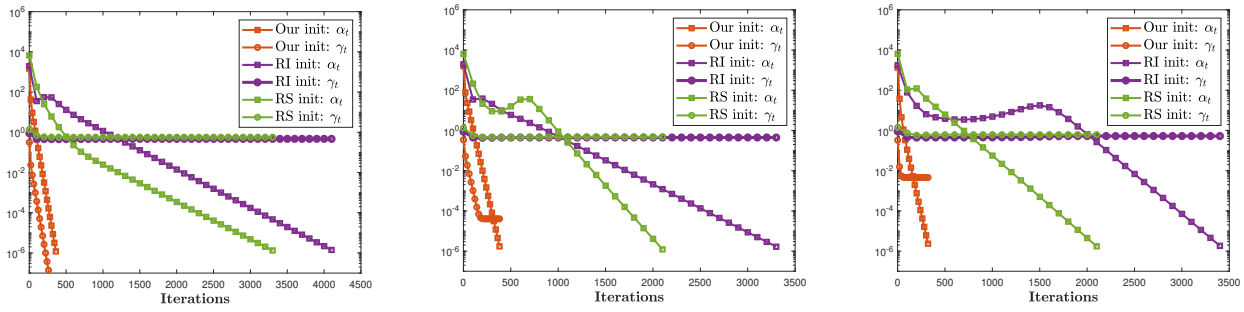
4.2. Convergence Behavior and Recovery Performance

In this subsection, we study the convergence behavior and recovery performance of GD in Algorithm 1 for solving Problem (3) in the ReLU sampling regime. To measure the convergence and recovery performance of the studied algorithms, we employ the following metrics at the t-th iteration:

$$\text{Norm of gradient: } \alpha_t = \|\nabla F(\boldsymbol{U}^{(t)})\|_F,$$
$$\text{Function value: } \beta_t = F(\boldsymbol{U}^{(t)}),$$
$$\text{Completion error: } \gamma_t = \|\boldsymbol{U}^{(t)}\boldsymbol{U}^{(t)^T} - \boldsymbol{M}\|_F/\|\boldsymbol{M}\|_F.$$

Obviously, if α_t is smaller, then $\boldsymbol{U}^{(t)}$ will be closer to a critical point. Moreover, if β_t and γ_t are smaller, then $\boldsymbol{U}^{(t)}$ will be closer to ground truth.

Our tailor-designed vs. random-imputation spectral initialization (RI) & random spectral (RS) initialization. To demonstrate the effectiveness of our tailor-designed initialization, we compare our proposed initialization with a random-imputation spectral (RI) and a random spectral (RS) initialization. Specifically, the RI initialization proceeds as follows: we generate a matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times r}$, whose entries are i.i.d. sampled from the standard normal distribution. Then, we construct $\boldsymbol{Q} = \boldsymbol{Y}\boldsymbol{Y}^T$ and generate a matrix $\bar{\boldsymbol{Q}}$ via setting $\bar{q}_{ij} = m_{ij}$ if $(i,j) \in \Omega$. Finally, we generate



(a) Noiseless case: $\sigma = 0$ (b) Noisy case: $\sigma = 10^{-4}$ (c) Noisy case: $\sigma = 10^{-2}$

Figure 5. **Convergence and recovery performance of GD for MC with ReLU sampling under different initialization schemes.**

Table 1. Comparison of the norm of gradient, function value, and completion error returned by GD with different initialization.

$\sigma = 0$	α_T	β_T	γ_T
Our init	$(9.7 \pm 0.16) \cdot 10^{-7}$	$(3.3 \pm 0.36) \cdot 10^{-14}$	$(7.6 \pm 0.7) \cdot 10^{-11}$
RI init	$(6.8 \pm 0.3) \cdot 10^{-3}$	$(7.9 \pm 2.1) \cdot 10^{3}$	0.5 ± 0.13
RS init	0.1 ± 0.45	$(7.1 \pm 3.8) \cdot 10^{3}$	0.45 ± 0.24
$\sigma = 10^{-4}$	α_T	β_T	γ_T
Our init	$(9.7 \pm 0.14) \cdot 10^{-7}$	$(1.8 \pm 0.03) \cdot 10^{-4}$	$(4.4 \pm 0.13) \cdot 10^{-5}$
RI init	$(0.39 \pm 1.2) \cdot 10^{-5}$	$(8.4 \pm 2.1) \cdot 10^{3}$	0.51 ± 0.13
RS init	$(1.5 \pm 6.5) \cdot 10^{-3}$	$(8.4 \pm 2.0) \cdot 10^{3}$	0.51 ± 0.13

$\boldsymbol{U}^{(0)} \in \mathcal{O}^{n \times r}$ using the truncated SVD to $\bar{\boldsymbol{Q}}$ as explained in Section 4.1. Moreover, the RS initialization proceeds exactly the same way as above, but without the imputation scheme $\bar{q}_{ij} = m_{ij}$ if $(i, j) \in \Omega$.

Experimental setup. In our experiments, we set $n = 200$ and $r = 5$. We generate data matrix $\boldsymbol{M}$ according to the model (1) with different noise level $\sigma \in \{0, 10^{-4}, 10^{-2}\}$ and sample the observed entries via ReLU sampling, e.g. (2). For each noise level, we generate 20 data matrices and run GD with our proposed, RI, and RS initialization on each data matrix, respectively. In each test, we terminate the algorithm when the Frobenius norm of the gradient at the T-th iteration is less than 10^{-6} or $T \geq 5000$.

Experimental results. To demonstrate the convergence and recovery performance of our proposed approach, we calculate and report the mean and standard deviation of the norm of gradient α_T, function value β_T, and completion error γ_T averaged over 20 runs in Table 1. An additional result for $\sigma = 10^{-2}$ can be found in Appendix C.1. In addition, we select one run from the 20 runs and plot norms of gradients $\{\alpha_t\}_{t \geq 0}$ and completion errors $\{\gamma_t\}_{t \geq 0}$ against the number of iterations in Figure 5. In the noiseless case, i.e., $\sigma = 0$, it is observed that GD with the tailor-designed initialization efficiently finds the ground truth at a linear rate, while GD with the RI or RS initialization converges a critical point that is not globally optimal. In the noisy case, i.e., $\sigma = 10^{-4}$, we observe that GD with the tailor-designed initialization converges to a critical point within an $O(\sigma)$-neighborhood of the ground truth, while GD with the RI or RS initialization converges a critical point that is significantly distant from the ground truth.

The results in Table 1 and Figure 5 support our theoretical results in Theorem 3.3 and Theorem 3.5. They empirically demonstrate that the optimal solutions to Problem (3) can recover the ground truth but this problem only has a local benign optimization landscape. Additionally, the comparison between our proposed initialization and the RI spectral initialization further highlights the effectiveness of our tailor-designed initialization.

4.3. Comparison to Existing Methods

In this subsection, we compare our proposed method with some state-of-the-art methods for solving MC in the ReLU sampling regime in terms of recovery performance and computational efficiency. This includes the alternating direction method of multipliers (ADMM) (Boyd et al., 2011) for the convex formulation of MC, scaled gradient descent (ScaledGD) for MC studied in (Tong et al., 2021), proximal alternating minimization (PAM) for solving the formulation proposed in (Saul, 2022), momentum proximal alternating minimization (MPAM) for solving nonlinear matrix decomposition studied in (Seraghiti et al., 2023), and Gaussian-Newton matrix recovery (GNMR) method for low-rank MC (Zilber & Nadler, 2022). To guarantee the performance of

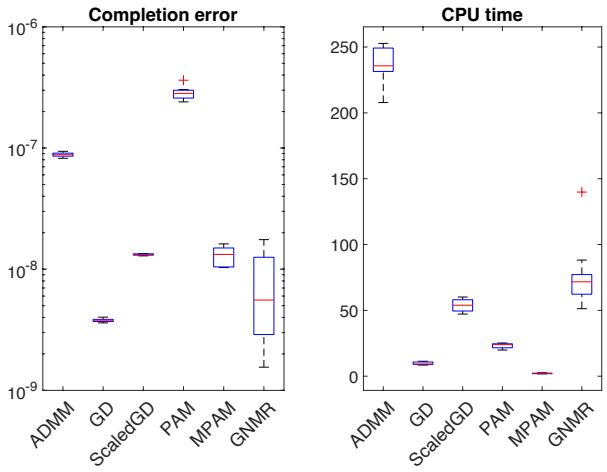


Figure 6. Comparison of completion error and CPU time for a variety of MC algorithms in the noiseless case.

ScaleGD, we employ our tailored-designed initialization scheme in Section 4.1 to initialize it. We refer the reader to Appendix C.2 for the problem formulations and implementation details.

Experimental setup. In our experiments, we set $n = 1000, r = 20$ and generate data matrix $\boldsymbol{M}$ according to the model (1) with different noise levels $\sigma \in \{0, 10^{-2}\}$. For each noise level, we generate 10 data matrices and run the MC algorithms on each data matrix. In each test, we terminate our algorithm when the Frobenius norm of the gradient is less than 10^{-4}. The termination criteria of other algorithms are specified in Appendix C.2.

Experimental results. To compare the recovery performance and computational efficiency of the tested methods, we calculate and report the completion error and CPU times averaged over 10 runs using box plots in Figure 6. See an additional figure for the noisy case in Appendix C.1. It is observed that our proposed method can achieve a comparable recovery performance to the other tested methods in both noiseless and noisy cases. Moreover, our proposed method is as fast as MPAM, slightly faster than PAM and ScaledGD, and substantially faster than ADMM and GNMR in terms of computational efficiency.

It is notable that all methods compared here have reasonable completion error on this ReLU MC problem. It was shown empirically in (Naik et al., 2022) that other methods, including direct nuclear norm minimization, do not work as well. It is of great interest to develop a broader theory to understand or clarify this discrepancy among algorithms for the ReLU sampling and more general entry-dependent MC problems.

4.4. Completion with various ranks

In Figure 7, we investigate completion accuracy in the noiseless setting when we increase the rank. In all cases, completion is successful well beyond our assumed rank bound of $r \leq \frac{1}{2} \log n$. Consider $n = 200$; even though $\log_2(200) \approx 7.6$, completion only begins to break down around $r = 45$. This behavior was also reported in (Naik et al., 2022). A deeper understanding of this fundamental limit is an exciting question for future work.

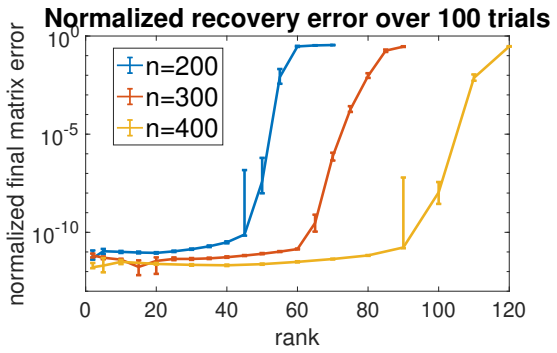


Figure 7. Completion in the noiseless case as rank varies, with dimension $n = 200, 300, 400$. Line shows the median error, and bars show the 25% and 75% error quantiles for 100 trials.

5. Conclusions

In this work, we studied the MC problem with ReLU sampling. Under mild assumptions on the underlying matrix, we showed that global optimal solutions of the low-rank formulation of MC recover the underlying matrix in both noiseless and noisy cases. Moreover, we also characterized the local optimization landscape, which is strongly convex on the quotient manifold in the neighborhood of global optimal solutions. Finally, we proposed a tailor-designed initialization for GD to optimize the studied formulation and conducted extensive experiments to showcase the potential of our proposed approach.

While our work fills a gap in the literature, there are still limitations that need to be addressed, starting with generalizing from SPSD matrices to rectangular matrices and considering unknown rank r. Weakening Assumption 3.6(d) is of great interest, as mentioned before. Additionally, we point out that our method and initialization have prior knowledge that may not be available in general, e.g., our initialization explicitly uses the ReLU sampling assumption. It would be interesting to empirically explore the sensitivity of MC algorithms to this coupling between sampling scheme and initialization. A final general future direction of study is to extend our landscape analysis to more general deterministic and entry-dependent sampling regimes.

Acknowledgements

The work of H.L. was supported in part by NSF China under Grant 12301403 and Young Elite Scientists Sponsorship Program by CAST 2023QNRC001. The work of

L.B. and P.W. was supported in part by ARO YIP award W911NF1910027, NSF CAREER award CCF-1845076, and DoE award DE-SC0022186. QQ acknowledges support from NSF CAREER CCF-2143904, NSF CCF-2212066, and NSF CCF-2212326.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Dahleh, M., Shah, D., and Shen, D. Causal matrix completion. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 3821–3826. PMLR, 2023.

Al-Homidan, S. and Wolkowicz, H. Approximate and exact completion problems for euclidean distance matrices using semidefinite programming. *Linear algebra and its applications*, 406:109–141, 2005.

Bhattacharya, S. and Chatterjee, S. Matrix completion with data-dependent missingness probabilities. *IEEE Transactions on Information Theory*, 2022.

Bishop, W. E. and Yu, B. M. Deterministic symmetric positive semidefinite matrix completion. *Advances in Neural Information Processing Systems*, 27, 2014.

Boumal, N. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Cai, H., Cai, J.-F., and You, J. Structured gradient descent for fast robust low-rank hankel matrix completion. *SIAM Journal on Scientific Computing*, 45(3):A1172–A1198, 2023.

Candes, E. and Recht, B. Exact matrix completion via convex optimization. *Communications of the ACM*, 55 (6):111–119, 2012.

Candes, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Chatterjee, S. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Chatterjee, S. A deterministic theory of low rank matrix completion. *IEEE Transactions on Information Theory*, 66(12):8046–8055, 2020.

Chen, Y. and Chi, Y. Spectral compressed sensing via structured matrix completion. In *International conference on machine learning*, pp. 414–422. PMLR, 2013.

Dokmanic, I., Parhizkar, R., Ranieri, J., and Vetterli, M. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32 (6):12–30, 2015.

Foucart, S., Needell, D., Pathak, R., Plan, Y., and Wootters, M. Weighted matrix completion from non-random, non-uniform sampling patterns. *IEEE Transactions on Information Theory*, 67(2):1264–1290, 2020.

Foygel, R., Shamir, O., Srebro, N., and Salakhutdinov, R. R. Learning with the weighted trace-norm under arbitrary sampling distributions. *Advances in neural information processing systems*, 24, 2011.

Ganti, R. S., Balzano, L., and Willett, R. Matrix completion under monotonic single index models. *Advances in neural information processing systems*, 28, 2015.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

Güler, O., Hoffman, A. J., and Rothblum, U. G. Approximations to solutions to systems of linear inequalities. *SIAM Journal on Matrix Analysis and Applications*, 16 (2):688–696, 1995.

Hernández-Lobato, J. M., Houlsby, N., and Ghahramani, Z. Probabilistic matrix factorization with non-random missing data. In *International Conference on Machine Learning*, pp. 1512–1520. PMLR, 2014.

Hoffman, A. J. On approximate solutions of systems of linear inequalities. In *Selected Papers Of Alan J Hoffman: With Commentary*, pp. 174–176. World Scientific, 2003.

Hosseini, S. M. and Sebt, M. A. Array interpolation using covariance matrix completion of minimum-size virtual array. *IEEE Signal Processing Letters*, 24(7):1063–1067, 2017.

Hu, Y., Koren, Y., and Volinsky, C. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pp. 263–272. Ieee, 2008.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th international conference on computer vision*, pp. 2146–2153. IEEE, 2009.

Ji, H., Liu, C., Shen, Z., and Xu, Y. Robust video denoising using low rank matrix completion. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 1791–1798. IEEE, 2010.

Jin, H., Ma, Y., and Jiang, F. Matrix completion with covariate information and informative missingness. *The Journal of Machine Learning Research*, 23(1):8115–8176, 2022.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

Király, F. J., Theran, L., and Tomioka, R. The algebraic combinatorial approach for low-rank matrix completion. *J. Mach. Learn. Res.*, 16(1):1391–1436, 2015.

Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37, 2009.

Laurent, M. Polynomial instances of the positive semidefinite and euclidean distance matrix completion problems. *SIAM Journal on Matrix Analysis and Applications*, 22 (3):874–894, 2001.

Lee, H., Mazumder, R., Song, Q., and Honorio, J. Matrix completion from general deterministic sampling patterns. *arXiv preprint arXiv:2306.02283*, 2023.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257. PMLR, 2016.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 3rd edition, 2019.

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

Liu, G., Liu, Q., Yuan, X.-T., and Wang, M. Matrix completion with deterministic sampling: Theories and methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):549–566, 2019.

Ma, C. and Zhang, C. Identifiable generative models for missing not at random data imputation. *Advances in Neural Information Processing Systems*, 34:27645–27658, 2021.

Ma, W. and Chen, G. H. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. *Advances in Neural Information Processing Systems*, 32, 2019.

Mazumdar, A. and Rawat, A. S. Representation learning and recovery in the ReLU model. *arXiv preprint arXiv:1803.04304*, 2018.

Naik, R., Trivedi, N., Tarzanagh, D. A., and Balzano, L. Truncated matrix completion-an empirical study. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 847–851. IEEE, 2022.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

Nguyen, L. T., Kim, J., and Shim, B. Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7: 94215–94237, 2019.

Ongie, G., Pimentel-Alarcon, D., Nowak, R., Willett, R., and Balzano, L. Tensor methods for nonlinear matrix completion. *SIAM journal on mathematics of data science*, 2020.

Pena, J., Vera, J. C., and Zuluaga, L. F. New characterizations of hoffman constants for systems of linear constraints. *Mathematical Programming*, 187:79–109, 2021.

Pimentel-Alarcón, D. L., Boston, N., and Nowak, R. D. A characterization of deterministic sampling patterns for low-rank matrix completion. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):623–636, 2016.

Recht, B. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

Saul, L. K. A nonlinear matrix decomposition for mining the zeros of sparse data. *SIAM Journal on Mathematics of Data Science*, 4(2):431–463, 2022.

Sengupta, N., Udell, M., Srebro, N., and Evans, J. Sparse data reconstruction, missing value and multiple imputation through matrix factorization. *Sociological Methodology*, 53(1):72–114, 2023.

Seraghiti, G., Awari, A., Vandaele, A., Porcelli, M., and Gillis, N. Accelerated algorithms for nonlinear matrix decomposition with the ReLU function. *arXiv preprint arXiv:2305.08687*, 2023.

Shamir, O. and Shalev-Shwartz, S. Matrix completion with the trace norm: Learning, bounding, and transducing. *The Journal of Machine Learning Research*, 15(1):3401–3423, 2014.

Shapiro, A., Xie, Y., and Zhang, R. Matrix completion with deterministic pattern: A geometric perspective. *IEEE Transactions on Signal Processing*, 67(4):1088–1103, 2018.

Singer, A. and Cucuringu, M. Uniqueness of low-rank matrix completion by rigidity theory. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1621–1641, 2010.

Soltanolkotabi, M. Learning ReLUs via gradient descent. *Advances in neural information processing systems*, 30, 2017.

Sportisse, A., Boyer, C., and Josse, J. Estimation and imputation in probabilistic principal component analysis with missing not at random data. *Advances in Neural Information Processing Systems*, 33:7067–7077, 2020a.

Sportisse, A., Boyer, C., and Josse, J. Imputation and low-rank estimation with missing not at random data. *Statistics and Computing*, 30(6):1629–1643, 2020b.

Sun, R. and Luo, Z.-Q. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Szarek, S. J. Metric entropy of homogeneous spaces. *arXiv preprint math/9701213*, 1997.

Tong, T., Ma, C., and Chi, Y. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *The Journal of Machine Learning Research*, 22(1):6639–6701, 2021.

Usevich, K. and Comon, P. Hankel low-rank matrix completion: Performance of the nuclear norm relaxation. *IEEE Journal of Selected Topics in Signal Processing*, 10(4): 637–646, 2016.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wersing, H., Beyn, W.-J., and Ritter, H. Dynamical stability conditions for recurrent neural networks with unsaturating piecewise linear transfer functions. *Neural Computation*, 13(8):1811–1825, 2001.

Yang, C., Ding, L., Wu, Z., and Udell, M. Tenips: Inverse propensity sampling for tensor completion. In *International Conference on Artificial Intelligence and Statistics*, pp. 3160–3168. PMLR, 2021.

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1025–1033, 2013.

Zilber, P. and Nadler, B. GNMR: a provable one-line algorithm for low rank matrix recovery. *SIAM Journal on Mathematics of Data Science*, 4(2):909–934, 2022.

Supplementary Material

The appendix is organized as follows: Appendix A contains all the proofs for the main results. Specifically, in Section A.1, we focus on the noiseless ReLU case and provide proofs for all the theorems presented in Section 3.1. In Section A.2, we shift our attention to the noisy general case and provide proofs for the theorems found in Section 3.2. In Section A.3, we concentrate on the Gaussian and ReLU sampling settings and present all the required proofs for Section 3.3. Appendix B provides the detailed related work. Appendix C provides detailed setups of our experiments and includes additional numerical results to supplement Section 4.

A. Proofs of Main Results

Before we proceed, let us recall some definitions and notions. Based on the block partition in (4), we denote by $\boldsymbol{M}^{(i,j)} \in \mathbb{R}^{n_i \times n_j}$ the (i,j)-th block of $\boldsymbol{M}$, i.e., $\boldsymbol{M}^{(i,j)} = \boldsymbol{U}_i^\star \boldsymbol{U}_j^{\star T}$ for any pair of indices $(i,j) \in [2^r] \times [2^r]$. Moreover, we denote the set of observations in the (i,j) block by $\Omega_{i,j} = \{(k,l) \in [n_i] \times [n_j] : m_{kl}^{(i,j)} \geq 0\}$ where $m_{kl}^{(i,j)}$ is the (k,l)-th entry of $\boldsymbol{M}^{(i,j)}$. Let a column vector $\boldsymbol{u}_{i,k}^\star \in \mathbb{R}^r$ denote the k-th row of $\boldsymbol{U}_i^\star$ for each $k \in [n_i]$. Then, we have $m_{kl}^{(i,j)} = \boldsymbol{u}_{i,k}^{\star T} \boldsymbol{u}_{j,l}^\star$. Since $\boldsymbol{u}_{i,k}^{\star T} \boldsymbol{u}_{i,l}^\star \geq 0$ always holds for all $k,l \in [n_i]$ due to the fact that the rows of $\boldsymbol{U}_i^\star$ have the same sign, we have $\Omega_{i,i} = [n_i] \times [n_i]$ for all $i \in [2^r]$ under the ReLU sampling (2).

A.1. Proofs for Section 3.1: Noiseless Case and ReLU Sampling

Firstly, we need to show that the horizontal space $\mathcal{S}_{\boldsymbol{U}}^\perp$, which is the orthogonal complement of $\mathcal{S}_{\boldsymbol{U}}$ defined in (6), is given by (7). Indeed, for arbitrary $\boldsymbol{D} \in \mathcal{S}_{\boldsymbol{U}}^\perp$, due to the orthogonality, we obtain that $\langle \boldsymbol{D}, \boldsymbol{U}\boldsymbol{R} \rangle = \langle \boldsymbol{U}^T \boldsymbol{D}, \boldsymbol{R} \rangle = 0$ holds for all skew-symmetric matrix $\boldsymbol{R}$. This holds if and only if $\boldsymbol{U}^T \boldsymbol{D}$ is symmetric. Consequently, $\mathcal{S}_{\boldsymbol{U}}^\perp$ is given by

$$\mathcal{S}_{\boldsymbol{U}}^\perp = \left\{ \boldsymbol{D} \in \mathbb{R}^{n \times r} \mid \boldsymbol{U}^T \boldsymbol{D} = \boldsymbol{D}^T \boldsymbol{U} \right\}. \tag{14}$$

A.1.1. Proof of Theorem 3.3

Proof of Theorem 3.3. Let $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ be an arbitrary optimal solution to Problem (3). According to (1) and $\boldsymbol{\Delta} = \boldsymbol{0}$, we have $\boldsymbol{M} = \boldsymbol{U}^\star \boldsymbol{U}^{\star T}$. This, together with the optimality of Problem (3), implies that $(\boldsymbol{U}\boldsymbol{U}^T)_\Omega = (\boldsymbol{U}^\star \boldsymbol{U}^{\star T})_\Omega$. Let $\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_1^T & \boldsymbol{U}_2^T & \cdots & \boldsymbol{U}_{2^r}^T \end{bmatrix}^T$ be the same partition as that in (4). Since all diagonal blocks are fully observed, we have $\boldsymbol{U}_i \boldsymbol{U}_i^T = \boldsymbol{U}_i^\star \boldsymbol{U}_i^{\star T}$. This, together with Assumption 3.1 and Lemma 3.4, yields that there exists $\boldsymbol{Q}_i \in \mathcal{O}^r$ such that $\boldsymbol{U}_i = \boldsymbol{U}_i^\star \boldsymbol{Q}_i$ for all $i \in [2^r]$.

Next, we show that the above $\boldsymbol{Q}_i$ for all $i \in [2^r]$ are the same. For each tri-diagonal block, it follows from $(\boldsymbol{U}\boldsymbol{U}^T)_\Omega = (\boldsymbol{U}^\star \boldsymbol{U}^{\star T})_\Omega$ that $(\boldsymbol{U}_{i+1} \boldsymbol{U}_i^T)_{\Omega_{i+1,i}} = (\boldsymbol{U}_{i+1}^\star \boldsymbol{U}_i^{\star T})_{\Omega_{i+1,i}}$ for all $i \in [2^r - 1]$. By substituting the equality $\boldsymbol{U}_i = \boldsymbol{U}_i^\star \boldsymbol{Q}_i$ into it, we get $(\boldsymbol{U}_{i+1}^\star \boldsymbol{Q}_{i+1} \boldsymbol{Q}_i^T \boldsymbol{U}_i^{\star T})_{\Omega_{i+1,i}} = (\boldsymbol{U}_{i+1}^\star \boldsymbol{U}_i^{\star T})_{\Omega_{i+1,i}}$. We write its entry-wise form as follows:

$$\boldsymbol{u}_{i+1,k}^{\star T} \boldsymbol{Q}_{i+1} \boldsymbol{Q}_i^T \boldsymbol{u}_{i,l}^\star = \boldsymbol{u}_{i+1,k}^{\star T} \boldsymbol{u}_{i,l}^\star, \quad \forall (k,l) \in \Omega_{i+1,i}. \tag{15}$$

According to Assumption 3.2, $\{\boldsymbol{u}_{i+1,k}^\star \boldsymbol{u}_{i,l}^{\star T} : (k,l) \in \Omega_{i+1,i}\}$ spans the whole r-by-r matrix space, so (15) holds if and only if $\boldsymbol{Q}_{i+1} \boldsymbol{Q}_i^T = \boldsymbol{I}_r$. Therefore, $\boldsymbol{Q}_i = \boldsymbol{Q}_j$ for all $i \neq j$. This yields $\boldsymbol{U} = \boldsymbol{U}^\star \boldsymbol{Q}$ with $\boldsymbol{Q} \in \mathcal{O}^r$ for any optimal solution $\boldsymbol{U}$. Then, we complete the proof. □

A.1.2. Proof of Theorem 3.5

Before we proceed, let us compute the gradient and Hessian of $F(\boldsymbol{U})$. Obviously, the gradient is

$$\nabla F(\boldsymbol{U}) = (\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{M})_\Omega \boldsymbol{U}. \tag{16}$$

Given a direction $\boldsymbol{D} \in \mathbb{R}^{n \times r}$, we have

$$\begin{aligned} \nabla F(\boldsymbol{U} + t\boldsymbol{D}) &= (\boldsymbol{U}\boldsymbol{U}^T + t\boldsymbol{U}\boldsymbol{D}^T + t\boldsymbol{D}\boldsymbol{U}^T + t^2 \boldsymbol{D}\boldsymbol{D}^T - \boldsymbol{M})_\Omega (\boldsymbol{U} + t\boldsymbol{D}) \\ &= \nabla F(\boldsymbol{U}) + t(\boldsymbol{U}\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{U}^T)_\Omega \boldsymbol{U} + t^2 (\boldsymbol{D}\boldsymbol{D}^T)_\Omega \boldsymbol{U} + t(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{M})_\Omega \boldsymbol{D} \\ &\quad t^2 (\boldsymbol{U}\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{U}^T)_\Omega \boldsymbol{D} + t^3 (\boldsymbol{D}\boldsymbol{D}^T)_\Omega \boldsymbol{D}. \end{aligned}$$

Then, we compute its bilinear form of the Hessian along a direction $\boldsymbol{D} \in \mathbb{R}^{n\times r}$ as follows:

$$\begin{aligned}\nabla^2 F(\boldsymbol{U})[\boldsymbol{D},\boldsymbol{D}] &= \left\langle \boldsymbol{D}, \lim_{t\to 0}\frac{\nabla F(\boldsymbol{U}+t\boldsymbol{D})-\nabla F(\boldsymbol{U})}{t}\right\rangle = \langle \boldsymbol{D}, (\boldsymbol{U}\boldsymbol{U}^T-\boldsymbol{M})_\Omega \boldsymbol{D} + (\boldsymbol{U}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^T)_\Omega \boldsymbol{U}\rangle \\ &= \langle \boldsymbol{D}\boldsymbol{D}^T, (\boldsymbol{U}\boldsymbol{U}^T-\boldsymbol{M})_\Omega\rangle + \langle \boldsymbol{D}\boldsymbol{U}^T, (\boldsymbol{U}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^T)_\Omega\rangle \\ &= \langle \boldsymbol{D}\boldsymbol{D}^T, (\boldsymbol{U}\boldsymbol{U}^T-\boldsymbol{M})_\Omega\rangle + \frac{1}{2}\|(\boldsymbol{U}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^T)_\Omega\|_F^2,\end{aligned} \tag{17}$$

where the last inequality holds because $\langle \boldsymbol{D}\boldsymbol{U}^T, (\boldsymbol{U}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^T)_\Omega\rangle = \langle \boldsymbol{U}\boldsymbol{D}^T, (\boldsymbol{U}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^T)_\Omega\rangle$ and

$$\langle \boldsymbol{D}\boldsymbol{U}^T, (\boldsymbol{U}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^T)_\Omega\rangle = \frac{1}{2}\langle \boldsymbol{U}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^T, (\boldsymbol{U}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^T)_\Omega\rangle = \frac{1}{2}\langle (\boldsymbol{U}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^T)_\Omega, (\boldsymbol{U}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^T)_\Omega\rangle.$$

Proof of Theorem 3.5. Using (17) and $\boldsymbol{M} = \boldsymbol{U}^\star\boldsymbol{U}^{\star T}$, we have

$$\nabla^2 F(\boldsymbol{U}^\star)[\boldsymbol{D},\boldsymbol{D}] = \frac{1}{2}\|(\boldsymbol{U}^\star\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^{\star T})_\Omega\|_F^2 \geq 0. \tag{18}$$

This implies that the Hessian of F at $\boldsymbol{U}^\star$ is semi-definite positive.

Next, we consider these directions $\boldsymbol{D}$ such that $\nabla^2 F(\boldsymbol{U}^\star)[\boldsymbol{D},\boldsymbol{D}] = 0$. We aim to show

$$\boldsymbol{D} = \boldsymbol{U}^\star\boldsymbol{V}, \text{ where } \boldsymbol{V}+\boldsymbol{V}^T = \boldsymbol{0}. \tag{19}$$

First, for each $i \in [2^r]$, since the i-th diagonal part is fully observed, i.e., $\Omega_{i,i} = [n_i]\times[n_i]$, we have

$$\boldsymbol{U}_i^\star\boldsymbol{D}_i^T + \boldsymbol{D}_i\boldsymbol{U}_i^{\star T} = \boldsymbol{0}. \tag{20}$$

It follows from Assumption 3.1 that $\mathrm{rank}(\boldsymbol{U}_i^\star) = r$, and thus its pseudo-inverse is $\boldsymbol{U}_i^{\star\dagger} = (\boldsymbol{U}_i^{\star T}\boldsymbol{U}_i^\star)^{-1}\boldsymbol{U}_i^{\star T} \in \mathbb{R}^{r\times n_i}$. By multiplying $\boldsymbol{U}_i^{\star\dagger T}$ on the right hand side of (20), we get $\boldsymbol{D}_i = -\boldsymbol{U}_i^\star\boldsymbol{D}_i^T\boldsymbol{U}_i^{\star\dagger T} = \boldsymbol{U}_i^\star\boldsymbol{V}_i$, where $\boldsymbol{V}_i := -\boldsymbol{D}_i^T\boldsymbol{U}_i^{\star\dagger T}$. Moreover, by multiplying $\boldsymbol{U}_i^{\star\dagger}$ and $\boldsymbol{U}_i^{\star\dagger T}$ on the left and right sides of (20), we get $\boldsymbol{D}_i^T\boldsymbol{U}_i^{\star\dagger T} + \boldsymbol{U}_i^{\star\dagger}\boldsymbol{D}_i = \boldsymbol{0}$. This implies $\boldsymbol{V}_i + \boldsymbol{V}_i^T = \boldsymbol{0}$, which means $\boldsymbol{V}_i$ is a skew-symmetric matrix. Now, let us consider the $(i+1,i)$-th block of $(\boldsymbol{U}^\star\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{U}^{\star T})_\Omega$. According to (20) and $\boldsymbol{D}_i = \boldsymbol{U}_i^\star\boldsymbol{V}_i$ for all $i \in [2^r]$, we have

$$\boldsymbol{0} = \left(\boldsymbol{U}_{i+1}^\star\boldsymbol{D}_i^T + \boldsymbol{D}_{i+1}\boldsymbol{U}_i^{\star T}\right)_{\Omega_{i+1,i}} = \left(\boldsymbol{U}_{i+1}^\star\boldsymbol{V}_i^T\boldsymbol{U}_i^{\star T} + \boldsymbol{U}_{i+1}^\star\boldsymbol{V}_{i+1}\boldsymbol{U}_i^{\star T}\right)_{\Omega_{i+1,i}} = \left(\boldsymbol{U}_{i+1}^\star\left(\boldsymbol{V}_i^T+\boldsymbol{V}_{i+1}\right)\boldsymbol{U}_i^{\star T}\right)_{\Omega_{i+1,i}}.$$

Using the same argument in (15), we have $\boldsymbol{V}_i^T + \boldsymbol{V}_{i+1} = \boldsymbol{0}$. Thus, we have $\boldsymbol{V}_i = \boldsymbol{V}_{i+1}$ for all $i \in [2^r-1]$, and thus $\boldsymbol{V}_i$ for all $i \in [2^r]$ are the same. Consequently, the space spanned by all the directions $\boldsymbol{D}$ satisfying $\nabla F(\boldsymbol{U}^\star)[\boldsymbol{D},\boldsymbol{D}] = \boldsymbol{0}$ is exactly the vertical space $\mathcal{S}_{\boldsymbol{U}^\star}$ defined in (6), i.e.,

$$\mathcal{S}_{\boldsymbol{U}^\star} = \left\{\boldsymbol{D} \in \mathbb{R}^{n\times r} : \nabla F(\boldsymbol{U}^\star)[\boldsymbol{D},\boldsymbol{D}] = 0\right\}.$$

Define the constant

$$\gamma := \min_{\boldsymbol{D}\in\mathcal{S}_{\boldsymbol{U}^\star}^\perp, \|\boldsymbol{D}\|_F=1} \|(\boldsymbol{U}^\star\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^{\star T})_\Omega\|_F^2 > 0 \tag{21}$$

We claim that $\gamma > 0$. Now, we prove this claim by contradiction. Indeed, suppose that $\gamma = 0$. This implies that there exists a $\boldsymbol{D} \in \mathcal{S}_{\boldsymbol{U}^\star}^\perp$ with $\|\boldsymbol{D}\|_F = 1$ such that $\|(\boldsymbol{U}^\star\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{U}^{\star T})_\Omega\|_F^2 = 0$. This, together with (19), yields $\boldsymbol{D} \in \mathcal{S}_{\boldsymbol{U}^\star}$ and thus $\boldsymbol{D} \in \mathcal{S}_{\boldsymbol{U}^\star} \cap \mathcal{S}_{\boldsymbol{U}^\star}^\perp = \{\boldsymbol{0}\}$, which contradicts the fact that $\|\boldsymbol{D}\|_F = 1$. Then, for any $\boldsymbol{D} \in \mathcal{S}_{\boldsymbol{U}^\star}^\perp$, (18) shows that

$$\nabla^2 F(\boldsymbol{U}^\star)[\boldsymbol{D},\boldsymbol{D}] = \frac{1}{2}\|(\boldsymbol{U}^\star\boldsymbol{D}_{\mathcal{S}_{\boldsymbol{U}^\star}^\perp}^T + \boldsymbol{D}_{\mathcal{S}_{\boldsymbol{U}^\star}^\perp}\boldsymbol{U}^{\star T})_\Omega\|_F^2 \geq \frac{\gamma}{2}\|\boldsymbol{D}_{\mathcal{S}_{\boldsymbol{U}^\star}^\perp}\|_F^2 = \frac{\gamma}{2}\|\boldsymbol{D}\|_F^2. \tag{22}$$

where $\boldsymbol{D}_{\mathcal{S}_{\boldsymbol{U}^\star}^\perp} = \mathrm{Proj}_{\mathcal{S}_{\boldsymbol{U}^\star}^\perp}(\boldsymbol{D})$ is the projection of $\boldsymbol{D}$ on $\mathcal{S}_{\boldsymbol{U}^\star}^\perp$. Finally, substituting $\boldsymbol{D} = \mathrm{lift}_{\boldsymbol{U}^\star}(\boldsymbol{\xi}) \in \mathcal{S}_{\boldsymbol{U}^\star}^\perp$ into the above inequality for any $\boldsymbol{\xi} \in \mathrm{T}_{[\boldsymbol{U}^\star]}\mathcal{M}$, together with (8), yields the desired result. □

From the geodesic strong convexity on $\mathcal{M}$ to the strong convexity in Euclidean space. For any $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times r}$, let $\boldsymbol{U}^T\boldsymbol{U}^\star = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^T$ be an SVD of $\boldsymbol{U}^T\boldsymbol{U}^\star$, where $\boldsymbol{P}, \boldsymbol{Q} \in \mathcal{O}^r$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$. Moreover, let $\widetilde{\boldsymbol{U}} := \boldsymbol{U}\boldsymbol{P}\boldsymbol{Q}^T$ and $\boldsymbol{D} := \widetilde{\boldsymbol{U}} - \boldsymbol{U}^\star$, then we must have $\boldsymbol{D} \in \mathcal{S}^{\perp}_{\boldsymbol{U}^\star}$. This is because, for any $\boldsymbol{E} \in \mathcal{S}_{\boldsymbol{U}^\star} = \{\boldsymbol{U}^\star\boldsymbol{V} : \boldsymbol{V} + \boldsymbol{V}^T = 0\}$, we have

$$\langle \boldsymbol{E}, \widetilde{\boldsymbol{U}} - \boldsymbol{U}^\star\rangle = \langle \boldsymbol{U}^\star\boldsymbol{V}, \boldsymbol{U}\boldsymbol{P}\boldsymbol{Q}^T - \boldsymbol{U}^\star\rangle = \langle \boldsymbol{V}, \boldsymbol{U}^{\star T}\boldsymbol{U}\boldsymbol{P}\boldsymbol{Q}^T - \boldsymbol{U}^{\star T}\boldsymbol{U}^\star\rangle = \langle \boldsymbol{V}, \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^T - \boldsymbol{U}^{\star T}\boldsymbol{U}^\star\rangle = \boldsymbol{0}, \tag{23}$$

where the last equality holds because $\boldsymbol{V}$ is a skew-symmetric matrix and $\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^T - \boldsymbol{U}^{\star T}\boldsymbol{U}^\star$ is symmetric. Therefore, (22) implies that

$$\nabla^2 F(\boldsymbol{U}^\star)[\widetilde{\boldsymbol{U}} - \boldsymbol{U}^\star, \widetilde{\boldsymbol{U}} - \boldsymbol{U}^\star] \geq \frac{\gamma}{2}\|\widetilde{\boldsymbol{U}} - \boldsymbol{U}^\star\|_F^2. \tag{24}$$

A.1.3. Proof of Lemma 3.4

Proof of Lemma 3.4. According to $r = \mathrm{rank}(\boldsymbol{U})$, we obtain that the pseudo-inverse of $\boldsymbol{U}$ is $\boldsymbol{U}^\dagger := (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T \in \mathbb{R}^{r\times n}$. Then, one can verify $\|\boldsymbol{U}^\dagger\| \leq 1/\sigma_{\min}(\boldsymbol{U})$. Next, let $\boldsymbol{\Psi} := \boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T$. Then, we compute

$$\boldsymbol{\Psi}\boldsymbol{U}^{\dagger T} = \boldsymbol{U} - \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{U}^{\dagger T}, \quad \boldsymbol{U}^\dagger\boldsymbol{\Psi}\boldsymbol{U}^{\dagger T} = \boldsymbol{I}_r - \boldsymbol{U}^\dagger\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{U}^{\dagger T}.$$

Note that

$$\|\boldsymbol{U}^\dagger\boldsymbol{\Psi}\boldsymbol{U}^{\dagger T}\|_F \leq \|\boldsymbol{U}^\dagger\|^2\|\boldsymbol{\Psi}\|_F \leq \frac{\|\boldsymbol{\Psi}\|_F}{\sigma^2_{\min}(\boldsymbol{U})}. \tag{25}$$

Let $\boldsymbol{V}^T\boldsymbol{U}^{\dagger T} = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{H}^T$ be a singular value decomposition (SVD) of $\boldsymbol{V}^T\boldsymbol{U}^{\dagger T} \in \mathbb{R}^{r\times r}$, where $\boldsymbol{P}, \boldsymbol{H} \in \mathcal{O}^r$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$. Then, we have

$$\|\boldsymbol{I}_r - \boldsymbol{\Sigma}^2\|_F = \|\boldsymbol{I}_r - \boldsymbol{H}\boldsymbol{\Sigma}^2\boldsymbol{H}^T\|_F = \|\boldsymbol{I}_r - \boldsymbol{U}^\dagger\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{U}^{\dagger T}\|_F = \|\boldsymbol{U}^\dagger\boldsymbol{\Psi}\boldsymbol{U}^{\dagger T}\|_F \leq \frac{\|\boldsymbol{\Psi}\|_F}{\sigma^2_{\min}(\boldsymbol{U})},$$

where the inequality follows from (25). This implies

$$\|\boldsymbol{\Sigma} - \boldsymbol{I}_r\|_F = \|(\boldsymbol{\Sigma} + \boldsymbol{I}_r)^{-1}(\boldsymbol{\Sigma}^2 - \boldsymbol{I}_r)\|_F \leq \|\boldsymbol{\Sigma}^2 - \boldsymbol{I}_r\|_F \leq \frac{\|\boldsymbol{\Psi}\|_F}{\sigma^2_{\min}(\boldsymbol{U})}. \tag{26}$$

Finally, by choosing $\boldsymbol{Q} = \boldsymbol{P}\boldsymbol{H}^T$, we get

$$\boldsymbol{\Psi}\boldsymbol{U}^{\dagger T} = \boldsymbol{U} - \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{U}^{\dagger T} = \boldsymbol{U} - \boldsymbol{V}\boldsymbol{Q} + \boldsymbol{V}\boldsymbol{P}(\boldsymbol{I}_r - \boldsymbol{\Sigma})\boldsymbol{H}^T.$$

Therefore, we obtain

$$\|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{Q}\|_F \leq \|\boldsymbol{\Psi}\boldsymbol{U}^{\dagger T}\|_F + \|\boldsymbol{V}\boldsymbol{P}(\boldsymbol{I}_r - \boldsymbol{\Sigma})\boldsymbol{H}^T\|_F \leq \frac{\|\boldsymbol{\Psi}\|_F}{\sigma_{\min}(\boldsymbol{U})} + \frac{\|\boldsymbol{V}\|\|\boldsymbol{\Psi}\|_F}{\sigma^2_{\min}(\boldsymbol{U})},$$

where the last inequality follows from $\|\boldsymbol{U}^\dagger\| \leq 1/\sigma_{\min}(\boldsymbol{U})$ and (26). □

Now, we remove the full rank assumption $\mathrm{rank}(\boldsymbol{U}) = r$ and prove the following more general lemma.

Lemma A.1. *For arbitrary $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times r}$ with $r \leq n$, there exists an orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}^r$ such that*

$$\|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{Q}\|_F \leq \max\left\{\frac{\sigma^2_{\min}(\boldsymbol{U}) + \|\boldsymbol{V}\|}{\sigma^2_{\min}(\boldsymbol{U})}, \frac{1}{\sigma_{\min}(\boldsymbol{V})}\right\}\|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T\|_F, \tag{27}$$

where $\sigma_{\min}(\boldsymbol{U})$ and $\sigma_{\min}(\boldsymbol{V})$ denote the smallest non-zero singular value of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively.

Proof. For ease of exposition, we denote the rank of $\boldsymbol{U}$ by $\hat{r} = \mathrm{rank}(\boldsymbol{U})$, where $\hat{r} \leq r$. Let $\boldsymbol{U} = \boldsymbol{P}_1\boldsymbol{\Sigma}_1\boldsymbol{Q}_1^T$ be a compact SVD of $\boldsymbol{U}$, where $\boldsymbol{P}_1 \in \mathcal{O}^{n\times\hat{r}}$, $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{\hat{r}\times\hat{r}}$, and $\boldsymbol{Q}_1 \in \mathcal{O}^{\hat{r}}$. Notice that

$$\begin{aligned}
\|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T\|_F^2 &= \|\boldsymbol{P}_1\boldsymbol{P}_1^T(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T) + (\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T)\|_F^2 \\
&= \|\boldsymbol{P}_1\boldsymbol{P}_1^T(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T)\|_F^2 + \|(\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T)\|_F^2 \\
&= \|\boldsymbol{P}_1^T(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T)\|_F^2 + \|(\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)\boldsymbol{V}\boldsymbol{V}^T\|_F^2,
\end{aligned}$$

where the last equality holds because $\boldsymbol{P}_1^T\boldsymbol{P}_1 = \boldsymbol{I}$ and $(\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)\boldsymbol{U} = \boldsymbol{0}$. Using the same argument, we have

$$\begin{aligned}\|\boldsymbol{P}_1^T(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T)\|_F^2 &= \|\boldsymbol{P}_1^T(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T)\boldsymbol{P}_1\boldsymbol{P}_1^T\|_F^2 + \|\boldsymbol{P}_1^T(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T)(\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)\|_F^2 \\ &= \|\boldsymbol{P}_1^T(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T)\boldsymbol{P}_1\|_F^2 + \|\boldsymbol{P}_1^T\boldsymbol{V}\boldsymbol{V}^T(\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)\|_F^2 \\ &\geq \|\boldsymbol{P}_1^T(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T)\boldsymbol{P}_1\|_F^2 = \|\boldsymbol{\Sigma}_1^2 - \boldsymbol{P}_1^T\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{P}_1\|_F^2,\end{aligned}$$

where the last equality holds because $\boldsymbol{U} = \boldsymbol{P}_1\boldsymbol{\Sigma}_1\boldsymbol{Q}_1^T$. Putting the above results together yields

$$\|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{V}\boldsymbol{V}^T\|_F^2 \geq \|\boldsymbol{\Sigma}_1^2 - \boldsymbol{P}_1^T\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{P}_1\|_F^2 + \|(\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)\boldsymbol{V}\boldsymbol{V}^T\|_F^2. \tag{28}$$

Similarly, we also have

$$\|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{Q}\|_F^2 = \|\boldsymbol{P}_1^T(\boldsymbol{U} - \boldsymbol{V}\boldsymbol{Q})\|_F^2 + \|(\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)\boldsymbol{V}\boldsymbol{Q}\|_F^2. \tag{29}$$

Let $\boldsymbol{P}_1^T\boldsymbol{V} = \boldsymbol{P}_2\boldsymbol{\Sigma}_2\boldsymbol{Q}_2^T$ be a compact SVD of $\boldsymbol{P}_1^T\boldsymbol{V}$, where $\boldsymbol{P}_2 \in \mathcal{O}^{n\times\hat{r}}$, $\boldsymbol{\Sigma}_2 \in \mathbb{R}^{\hat{r}\times\hat{r}}$, and $\boldsymbol{Q}_2 \in \mathcal{O}^{\hat{r}}$. By choosing $\boldsymbol{Q}$ such that $\boldsymbol{Q}_2^T\boldsymbol{Q} = \boldsymbol{Q}_1^T$, we have

$$\|\boldsymbol{P}_1^T(\boldsymbol{U} - \boldsymbol{V}\boldsymbol{Q})\|_F = \|\boldsymbol{\Sigma}_1\boldsymbol{Q}_1^T - \boldsymbol{P}_1^T\boldsymbol{V}\boldsymbol{Q}\|_F = \|\boldsymbol{\Sigma}_1 - \boldsymbol{P}_2\boldsymbol{\Sigma}_2\|_F \leq \frac{\sigma_{\min}^2(\boldsymbol{U}) + \|\boldsymbol{V}\|}{\sigma_{\min}^2(\boldsymbol{U})}\|\boldsymbol{\Sigma}_1^2 - \boldsymbol{P}_1^T\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{P}_1\|_F^2, \tag{30}$$

where the second equality follows from $\boldsymbol{P}_1^T\boldsymbol{V} = \boldsymbol{P}_2\boldsymbol{\Sigma}_2\boldsymbol{Q}_2^T$ and $\boldsymbol{Q}_2^T\boldsymbol{Q} = \boldsymbol{Q}_1^T$, and the inequality uses Lemma 3.4. Moreover, we have

$$\|(\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)\boldsymbol{V}\boldsymbol{Q}\|_F^2 = \|(\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)\boldsymbol{V}_1\|_F^2 \leq \frac{1}{\sigma_{\min}(\boldsymbol{V})}\|(\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{P}_1^T)\boldsymbol{V}_1\boldsymbol{V}_1^T\|_F^2.$$

This, together with (28), (29), and (30), implies (27). □

A.2. Proofs for Section 3.2: Noisy Case with General Sampling

A.2.1. Proof of Theorem 3.8

Proof of Theorem 3.8. Using the global optimality of $\boldsymbol{U}$, (1), and (3), we have

$$F(\boldsymbol{U}) \leq F(\boldsymbol{U}^\star) = \frac{1}{4}\|\boldsymbol{\Delta}_\Omega\|_F^2 \leq \frac{1}{4}\|\boldsymbol{\Delta}\|_F^2. \tag{31}$$

In addition, we have

$$F(\boldsymbol{U}) = \frac{1}{4}\|(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{M})_\Omega\|_F^2 \geq \frac{1}{8}\|(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}^\star\boldsymbol{U}^{\star T})_\Omega\|_F^2 - \frac{1}{4}\|\boldsymbol{\Delta}\|_F^2,$$

where the inequality follows from (1) and $\|\boldsymbol{A} - \boldsymbol{B}\|_F^2 \geq \|\boldsymbol{A}\|_F^2/2 - \|\boldsymbol{B}\|_F^2$ for all $\boldsymbol{A}, \boldsymbol{B}$. This, together with (31), implies

$$\|(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}^\star\boldsymbol{U}^{\star T})_\Omega\|_F^2 \leq 4\|\boldsymbol{\Delta}\|_F^2. \tag{32}$$

Now, we consider the diagonal blocks of $\boldsymbol{M}$ index by $\mathcal{I}_k \times \mathcal{I}_k$. For each $k \in [K]$, let $\boldsymbol{C}_k := \boldsymbol{U}_k\boldsymbol{U}_k^T - \boldsymbol{U}_k^\star\boldsymbol{U}_k^{\star T}$. According to condition (a) in Assumption 3.6, the block $\mathcal{I}_k \times \mathcal{I}_k$ is fully observed, and thus we have for each $k \in [K]$,

$$\|\boldsymbol{C}_k\|_F \leq \|(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}^\star\boldsymbol{U}^{\star T})_\Omega\|_F \leq 2\|\boldsymbol{\Delta}\|_F. \tag{33}$$

Recall that $\boldsymbol{U}_k^{\star\dagger} = (\boldsymbol{U}_k^{\star T}\boldsymbol{U}_k^\star)^{-1}\boldsymbol{U}_k^{\star T}$. Let $\boldsymbol{U}_k^{\star\dagger}\boldsymbol{U}_k = \boldsymbol{P}_k\boldsymbol{\Sigma}_k\boldsymbol{H}_k^T$ be an SVD of $\boldsymbol{U}_k^{\star\dagger}\boldsymbol{U}_k \in \mathbb{R}^{r\times r}$, where $\boldsymbol{P}_k, \boldsymbol{H}_k \in \mathcal{O}^r$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{r\times r}$, and $\boldsymbol{Q}_k := \boldsymbol{H}_k\boldsymbol{P}_k^T$. According to Lemma 3.4 and the proof, we have

$$\begin{aligned}\|\boldsymbol{U}_k\boldsymbol{Q}_k - \boldsymbol{U}_k^\star\|_F &\leq \frac{\sigma_{\min}(\boldsymbol{U}_k^\star) + \|\boldsymbol{U}_k\|}{\sigma_{\min}^2(\boldsymbol{U}_k^\star)}\|\boldsymbol{U}_k\boldsymbol{U}_k^T - \boldsymbol{U}_k^\star\boldsymbol{U}_k^{\star T}\|_F \\ &\leq \frac{2(\sqrt{\lambda} + \|\boldsymbol{U}_k\|)}{\lambda}\|\boldsymbol{\Delta}\|_F \leq \frac{2}{\lambda}\|\boldsymbol{\Delta}\|_F\left(\sqrt{\lambda} + 2\|\boldsymbol{\Delta}\|^{\frac{1}{2}} + \|\boldsymbol{U}_k^\star\|\right),\end{aligned} \tag{34}$$

where the second inequality follows from (33) and $\sigma_{\min}(\boldsymbol{U}_k^\star) \geq \sqrt{\lambda}$ for all $k \in [K]$ due to condition (b) in Assumption 3.6, and the last inequality uses

$$\|\boldsymbol{U}_k\| = \|\boldsymbol{U}_k\boldsymbol{U}_k^T\|^{\frac{1}{2}} = \|\boldsymbol{C}_k + \boldsymbol{U}_k^\star\boldsymbol{U}_k^{\star T}\|^{\frac{1}{2}} \leq \|\boldsymbol{C}_k\|^{\frac{1}{2}} + \|\boldsymbol{U}_k^\star\| \leq 2\|\boldsymbol{\Delta}\|^{\frac{1}{2}} + \|\boldsymbol{U}_k^\star\|.$$

For these off-diagonal blocks satisfying $\mathcal{I}_j \times \mathcal{I}_k \in \Omega$, let $\boldsymbol{C}_{j,k} := \boldsymbol{U}_j\boldsymbol{U}_k^T - \boldsymbol{U}_j^\star\boldsymbol{U}_k^{\star T}$ for all $j \neq k \in [K]$. Similar to (33), we have $\|\boldsymbol{C}_{j,k}\|_F \leq 2\|\boldsymbol{\Delta}\|_F$ and

$$\boldsymbol{U}_j^{\star\dagger}\boldsymbol{U}_j(\boldsymbol{U}_k^{\star\dagger}\boldsymbol{U}_k)^T = \boldsymbol{I}_r + \boldsymbol{U}_j^{\star\dagger}\boldsymbol{C}_{j,k}\boldsymbol{U}_k^{\star\dagger T}.$$

Substituting the SVDs $\boldsymbol{U}_j^{\star\dagger}\boldsymbol{U}_j = \boldsymbol{P}_j\boldsymbol{\Sigma}_j\boldsymbol{H}_j^T$ and $\boldsymbol{U}_k^{\star\dagger}\boldsymbol{U}_k = \boldsymbol{P}_k\boldsymbol{\Sigma}_k\boldsymbol{H}_k^T$ into the above equality yields

$$\begin{aligned}\boldsymbol{I}_r + \boldsymbol{U}_j^{\star\dagger}\boldsymbol{C}_{j,k}(\boldsymbol{U}_k^{\star\dagger})^T &= \boldsymbol{U}_j^{\star\dagger}\boldsymbol{U}_j(\boldsymbol{U}_k^{\star\dagger}\boldsymbol{U}_k)^T = \boldsymbol{P}_j\boldsymbol{\Sigma}_j\boldsymbol{H}_j^T\boldsymbol{H}_k\boldsymbol{\Sigma}_k\boldsymbol{P}_k^T \\ &= \boldsymbol{P}_j(\boldsymbol{\Sigma}_j - \boldsymbol{I}_r)\boldsymbol{H}_j^T\boldsymbol{H}_k(\boldsymbol{\Sigma}_k - \boldsymbol{I}_r)\boldsymbol{P}_k^T + \boldsymbol{P}_j(\boldsymbol{\Sigma}_j - \boldsymbol{I}_r)\boldsymbol{H}_j^T\boldsymbol{H}_k\boldsymbol{P}_k^T + \\ &\quad \boldsymbol{P}_j\boldsymbol{H}_j^T\boldsymbol{H}_k(\boldsymbol{\Sigma}_k - \boldsymbol{I}_r)\boldsymbol{P}_k^T + \boldsymbol{P}_j\boldsymbol{H}_j^T\boldsymbol{H}_k\boldsymbol{P}_k^T\end{aligned}$$

Note that $\boldsymbol{Q}_k = \boldsymbol{H}_k\boldsymbol{P}_k^T$ for all $k \in [K]$, and thus $\boldsymbol{P}_j\boldsymbol{H}_j^T\boldsymbol{H}_k\boldsymbol{P}_k^T = \boldsymbol{Q}_j^T\boldsymbol{Q}_k$. Then, we have

$$\begin{aligned}\|\boldsymbol{Q}_j - \boldsymbol{Q}_k\|_F &= \|\boldsymbol{Q}_j\boldsymbol{Q}_k^T - \boldsymbol{I}_r\|_F = \|\boldsymbol{P}_j\boldsymbol{H}_j^T\boldsymbol{H}_k\boldsymbol{P}_k^T - \boldsymbol{I}_r\|_F \\ &= \|\boldsymbol{U}_j^{\star\dagger}\boldsymbol{C}_{j,k}(\boldsymbol{U}_k^{\star\dagger})^T - \boldsymbol{P}_j(\boldsymbol{\Sigma}_j - \boldsymbol{I}_r)\boldsymbol{H}_j^T\boldsymbol{H}_k(\boldsymbol{\Sigma}_k - \boldsymbol{I}_r)\boldsymbol{P}_k^T - \boldsymbol{P}_j(\boldsymbol{\Sigma}_j - \boldsymbol{I}_r)\boldsymbol{H}_j^T\boldsymbol{H}_k\boldsymbol{P}_k^T - \\ &\quad \boldsymbol{P}_j\boldsymbol{H}_j^T\boldsymbol{H}_k(\boldsymbol{\Sigma}_k - \boldsymbol{I}_r)\boldsymbol{P}_k^T\|_F \\ &\leq \|\boldsymbol{\Sigma}_j - \boldsymbol{I}_r\|_F\|\boldsymbol{\Sigma}_k - \boldsymbol{I}_r\|_F + \|\boldsymbol{\Sigma}_j - \boldsymbol{I}_r\|_F + \|\boldsymbol{\Sigma}_k - \boldsymbol{I}_r\|_F + \|\boldsymbol{U}_j^{\star\dagger}\boldsymbol{C}_{j,k}\boldsymbol{U}_k^{\star\dagger T}\|_F \\ &\leq \frac{1}{\lambda^2}\|\boldsymbol{C}_j\|_F\|\boldsymbol{C}_k\|_F + \frac{1}{\lambda}\|\boldsymbol{C}_j\|_F + \frac{1}{\lambda}\|\boldsymbol{C}_k\|_F + \frac{1}{\lambda}\|\boldsymbol{C}_{j,k}\|_F \leq \frac{2}{\lambda}\|\boldsymbol{\Delta}\|_F\left(3 + \frac{2}{\lambda}\|\boldsymbol{\Delta}\|_F\right),\end{aligned} \tag{35}$$

where the second inequality follows from (26). For each $k \in [K]$, according to condition (d) in Assumption 3.6, there exists a path $k_0 \to k_1 \to \cdots \to k_s$ such that $k_0 = k$, $k_s = 1$ and $\mathcal{I}_{k_{j-1}} \times \mathcal{I}_{k_j} \subseteq \Omega$ for all $j \in [s]$. Therefore, we bound the term $\boldsymbol{Q}_k - \boldsymbol{Q}_1$ as follows:

$$\|\boldsymbol{Q}_k - \boldsymbol{Q}_1\|_F \leq \sum_{j\in[s]}\|\boldsymbol{Q}_{k_{j-1}} - \boldsymbol{Q}_{k_j}\|_F \leq \frac{2K}{\lambda}\|\boldsymbol{\Delta}\|_F\left(3 + \frac{2}{\lambda}\|\boldsymbol{\Delta}\|_F\right), \tag{36}$$

where the last inequality follows from (35) and $s \leq K$. Therefore, we obtain

$$\begin{aligned}\|\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{Q}_1^T\|_F^2 &= \sum_{k=1}^K\|\boldsymbol{U}_k - \boldsymbol{U}_k^\star\boldsymbol{Q}_1^T\|_F^2 \leq 2\sum_{k=1}^K\left(\|\boldsymbol{U}_k - \boldsymbol{U}_k^\star\boldsymbol{Q}_k^T\|_F^2 + \|\boldsymbol{U}_k^\star(\boldsymbol{Q}_k - \boldsymbol{Q}_1)^T\|_F^2\right) \\ &\leq \frac{8K}{\lambda^2}\|\boldsymbol{\Delta}\|_F^2(\sqrt{\lambda} + 2\|\boldsymbol{\Delta}\|^{\frac{1}{2}} + \|\boldsymbol{U}^\star\|)^2 + 2K^2\|\boldsymbol{U}^\star\|^2\left(\frac{2}{\lambda}\|\boldsymbol{\Delta}\|_F\left(3 + \frac{2}{\lambda}\|\boldsymbol{\Delta}\|_F\right)\right)^2 \\ &\leq \frac{8K}{\lambda^2}\|\boldsymbol{\Delta}\|_F^2\left(\left(\sqrt{\lambda} + 2\|\boldsymbol{\Delta}\|_F^{\frac{1}{2}} + \|\boldsymbol{U}^\star\|\right)^2 + K^2\|\boldsymbol{U}^\star\|^2\left(3 + \frac{2}{\lambda}\|\boldsymbol{\Delta}\|_F\right)^2\right)\end{aligned}$$

where the second inequality follows from (34) and (36), and the third inequality holds because $\|\boldsymbol{U}_k^\star(\boldsymbol{Q}_k - \boldsymbol{Q}_1)^T\|_F \leq \|\boldsymbol{U}_k^\star\|\|\boldsymbol{Q}_k - \boldsymbol{Q}_1\|_F$. As a result, we get

$$\|\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{Q}_1^T\|_F \leq \frac{c_0}{\lambda}\|\boldsymbol{\Delta}\|_F, \text{ where } c_0 = \sqrt{8K}\left(\sqrt{\lambda} + 2\|\boldsymbol{\Delta}\|_F^{\frac{1}{2}} + \|\boldsymbol{U}^\star\| + K\|\boldsymbol{U}^\star\|\left(3 + \frac{2}{\lambda}\|\boldsymbol{\Delta}\|_F\right)\right). \tag{37}$$

Finally, based on the fact that $\boldsymbol{U}^\star\boldsymbol{U}^{\star T} = \boldsymbol{U}^\star\boldsymbol{Q}_1^T(\boldsymbol{U}^\star\boldsymbol{Q}_1^T)^T$, we have

$$\begin{aligned}\|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}^\star\boldsymbol{U}^{\star T}\|_F =& \|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}^\star\boldsymbol{Q}_1^T(\boldsymbol{U}^\star\boldsymbol{Q}_1^T)^T\|_F \\ =& \|(\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{Q}_1^T)(\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{Q}_1^T)^T + (\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{Q}_1^T)(\boldsymbol{U}^\star\boldsymbol{Q}_1^T)^T + \boldsymbol{U}^\star\boldsymbol{Q}_1^T(\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{Q}_1^T)^T\|_F \\ \leq& \|\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{Q}_1^T\|_F\left(2\|\boldsymbol{U}^\star\| + \|\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{Q}_1^T\|\right) \\ \leq& \frac{c_0}{\lambda}\left(2\|\boldsymbol{U}^\star\| + \frac{c_0}{\lambda}\|\boldsymbol{\Delta}\|_F\right)\|\boldsymbol{\Delta}\|_F,\end{aligned} \tag{38}$$

where the first inequality holds because of the triangle inequality and the second inequality holds because of (37). We complete the proof by letting

$$c := \frac{c_0}{\lambda}\left(2\|\boldsymbol{U}^\star\|_2 + \frac{c_0}{\lambda}\|\boldsymbol{\Delta}\|_F\right)\|\boldsymbol{\Delta}\|_F. \tag{39}$$

□

A.2.2. PROOF OF THEOREM 3.9

Proof of Theorem 3.9. Recall that $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ is a global optimal solution of Problem (3). For ease of exposition, let $\widetilde{\boldsymbol{\Delta}} := (\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{M})_\Omega$. Obviously, we have $\|\widetilde{\boldsymbol{\Delta}}\|_F \le \|\boldsymbol{\Delta}\|_F$. This, together with (17), yields

$$\nabla^2 F(\boldsymbol{U})[\boldsymbol{D},\boldsymbol{D}] \ge \frac{1}{2}\|(\boldsymbol{U}\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{U}^T)_\Omega\|_F^2 - \|\boldsymbol{D}\boldsymbol{D}^T\|_F\|\widetilde{\boldsymbol{\Delta}}\|_F \ge \frac{1}{2}\|(\boldsymbol{U}\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{U}^T)_\Omega\|_F^2 - \|\boldsymbol{\Delta}\|_F\|\boldsymbol{D}\|_F^2. \tag{40}$$

Next, we consider some direction $\boldsymbol{D}$ such that $(\boldsymbol{U}\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{U}^T)_\Omega = \boldsymbol{0}$. We claim that $\boldsymbol{D} = \boldsymbol{U}\boldsymbol{V}$ for some skew-symmetric matrix $\boldsymbol{V} \in \mathbb{R}^{r\times r}$. Indeed, for each $i \in [2^r]$, since the i-th diagonal part is fully observed, i.e., $\Omega_{i,i} = [n_i]\times[n_i]$, we have

$$\boldsymbol{U}_i\boldsymbol{D}_i^T + \boldsymbol{D}_i\boldsymbol{U}_i^T = \boldsymbol{0}. \tag{41}$$

Using Weyl's inequality, we obtain

$$\sigma_r(\boldsymbol{U}_i) \ge \sigma_r(\boldsymbol{U}_i^\star) - \|\boldsymbol{U}_i\boldsymbol{Q}_i - \boldsymbol{U}_i^\star\|_F \ge \sqrt{\lambda} - \frac{2}{\lambda}\|\boldsymbol{\Delta}\|_F\left(\sqrt{\lambda}+\sqrt{\gamma}+\|\boldsymbol{U}^\star\|\right) \ge \frac{1}{2}\sqrt{\lambda},$$

where the second inequality follows from the assumption that $\boldsymbol{U}_i^{\star T}\boldsymbol{U}_i^\star \succeq \lambda\boldsymbol{I}_r$, (34), and $\|\boldsymbol{\Delta}\|_F < \gamma/8$ by (11), and the last inequality holds because of $\|\boldsymbol{\Delta}\|_F \le \lambda\sqrt{\lambda}/4(\sqrt{\lambda}+\sqrt{\gamma}+\|\boldsymbol{U}^\star\|)$ by (11). Therefore, the pseudo-inverse $\boldsymbol{U}_i^\dagger = (\boldsymbol{U}_i^T\boldsymbol{U}_i)^{-1}\boldsymbol{U}_i^T$ is well-defined for all $i \in [2^r]$. Multiplying $\boldsymbol{U}_i^{\dagger T}$ from the right-hand side of (41) yields $\boldsymbol{D}_i = -\boldsymbol{U}_i(\boldsymbol{U}_i^\dagger\boldsymbol{D}_i)^T = \boldsymbol{U}_i\boldsymbol{V}_i$, where $\boldsymbol{V}_i = -(\boldsymbol{U}_i^\dagger\boldsymbol{D}_i)^T$. Moreover, multiplying $\widehat{\boldsymbol{U}}_i^\dagger$ and $\widehat{\boldsymbol{U}}_i^{\dagger T}$ from the left and right hand sides of (41) yields $(\boldsymbol{U}_i^\dagger\boldsymbol{D}_i)^T + \boldsymbol{U}_i^\dagger\boldsymbol{D}_i = 0$. This implies $\boldsymbol{V}_i + \boldsymbol{V}_i^T = \boldsymbol{0}$, which means $\boldsymbol{V}_i$ is a skew matrix. Then, consider the (i,j)-th block of $(\boldsymbol{U}\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{U}^T)_\Omega$. Since $\boldsymbol{D}_i = \boldsymbol{U}_i\boldsymbol{V}_i$ for any $i \in [2^r]$, it follows from (41) that

$$\boldsymbol{0} = \boldsymbol{U}_j\boldsymbol{D}_i^T + \boldsymbol{D}_j\boldsymbol{U}_i^T = \boldsymbol{U}_j\boldsymbol{V}_i^T\boldsymbol{U}_i^T + \boldsymbol{U}_j\boldsymbol{V}_j\boldsymbol{U}_i^T = \boldsymbol{U}_{i+1}\left(\boldsymbol{V}_i^T + \boldsymbol{V}_{i+1}\right)\boldsymbol{U}_i^T.$$

Since $\boldsymbol{U}_i$ and $\boldsymbol{U}_{i+1}$ are full-rank, we obtain $\boldsymbol{V}_i^T + \boldsymbol{V}_{i+1} = \boldsymbol{0}$. Thus, we have $\boldsymbol{V}_i = \boldsymbol{V}_{i+1}$ for all $i \in [2^r - 1]$, and thus $\boldsymbol{V}_i$ for all $i \in [2^r]$ are the same.

For ease of exposition, let $\mathcal{S} = \mathcal{S}_{\boldsymbol{U}}$ and $\mathcal{S}^\perp = \mathcal{S}_{\boldsymbol{U}}^\perp$. For any $\boldsymbol{D} \in \mathbb{R}^{n\times r}$, we decompose it as

$$\boldsymbol{D} = \boldsymbol{D}_{\mathcal{S}} + \boldsymbol{D}_{\mathcal{S}^\perp}, \text{ where } \boldsymbol{D}_{\mathcal{S}} = \mathrm{proj}_{\mathcal{S}}(\boldsymbol{D}),\ \boldsymbol{D}_{\mathcal{S}^\perp} = \mathrm{proj}_{\mathcal{S}^\perp}(\boldsymbol{D}). \tag{42}$$

Define the constant

$$\widehat{\gamma} := \|(\boldsymbol{U}\widehat{\boldsymbol{E}}^T + \widehat{\boldsymbol{E}}\boldsymbol{U}^T)_\Omega\|_F^2, \text{ where } \widehat{\boldsymbol{E}} = \operatorname*{argmin}_{\boldsymbol{E}\in\mathcal{S}^\perp, \|\boldsymbol{E}\|_F=1} \|(\boldsymbol{U}\boldsymbol{E}^T + \boldsymbol{E}\boldsymbol{U}^T)_\Omega\|_F^2. \tag{43}$$

According to (40), we compute for any $\boldsymbol{D} \in \mathcal{S}^\perp$,

$$\nabla^2 F(\boldsymbol{U})[\boldsymbol{D},\boldsymbol{D}] \ge \frac{1}{2}\|(\boldsymbol{U}\boldsymbol{D}_{\mathcal{S}^\perp}^T + \boldsymbol{D}_{\mathcal{S}^\perp}\boldsymbol{U}^T)_\Omega\|_F^2 - \|\boldsymbol{\Delta}\|_F\|\boldsymbol{D}\|_F^2 \ge \frac{\widehat{\gamma}}{2}\|\boldsymbol{D}_{\mathcal{S}^\perp}\|_F^2 - \|\boldsymbol{\Delta}\|_F\|\boldsymbol{D}\|_F^2 = \left(\frac{\widehat{\gamma}}{2} - \|\boldsymbol{\Delta}\|_F\right)\|\boldsymbol{D}\|_F^2.$$

We complete the proof by letting $\boldsymbol{D} = \mathrm{lift}_{\boldsymbol{U}}(\boldsymbol{\xi})$ for any $\boldsymbol{\xi} \in \mathrm{T}_{[\boldsymbol{U}]}\mathcal{M}$ and showing that $\widehat{\gamma} \ge \gamma/4$.

Then, the rest of the proof is devoted to proving $\widehat{\gamma} \ge \gamma/4$. Without loss of generality, we assume that $\boldsymbol{Q}_1 = \boldsymbol{I}_r$ in (37) and let $\overline{\boldsymbol{U}} = \boldsymbol{U} - \boldsymbol{U}^\star$, then we have $\|\overline{\boldsymbol{U}}\|_F \le c_0\|\boldsymbol{\Delta}\|_F/\lambda \le \tilde{c}_0\|\boldsymbol{\Delta}\|_F/\lambda$ where

$$\begin{aligned}\tilde{c}_0 &= \sqrt{8K}\left(\sqrt{\lambda}+\sqrt{\gamma}+\|\boldsymbol{U}^\star\| + K\|\boldsymbol{U}^\star\|\left(3+\frac{\gamma}{4\lambda}\right)\right)\\ &\ge \sqrt{8K}\left(\sqrt{\lambda}+2\|\boldsymbol{\Delta}\|_F^{\frac{1}{2}}+\|\boldsymbol{U}^\star\| + K\|\boldsymbol{U}^\star\|\left(3+\frac{2}{\lambda}\|\boldsymbol{\Delta}\|_F\right)\right) = c_0,\end{aligned} \tag{44}$$

which holds because of the condition $\|\boldsymbol{\Delta}\|_F \leq \frac{\gamma}{8}$. Thus, we have

$$\sqrt{\widehat{\gamma}} \geq \|(\boldsymbol{U}^\star \widehat{\boldsymbol{E}}^T + \widehat{\boldsymbol{E}}\boldsymbol{U}^{\star T})_\Omega\|_F - \|(\overline{\boldsymbol{U}}\widehat{\boldsymbol{E}}^T + \widehat{\boldsymbol{E}}\overline{\boldsymbol{U}}^T)_\Omega\|_F \geq \|(\boldsymbol{U}^\star \widehat{\boldsymbol{E}}^T + \widehat{\boldsymbol{E}}\boldsymbol{U}^{\star T})_\Omega\|_F - \frac{2\tilde{c}_0}{\lambda}\|\boldsymbol{\Delta}\|_F, \tag{45}$$

where the first inequality holds because of the triangle inequality, and the second inequality follows from $\|\widehat{\boldsymbol{E}}\|_F = 1$ and $\|\overline{\boldsymbol{U}}\|_F \leq \tilde{c}_0\|\boldsymbol{\Delta}\|_F/\lambda$. Recall that γ is defined w.r.t. $\mathcal{S}^\perp_{\boldsymbol{U}^\star} = \{\boldsymbol{D} : \boldsymbol{U}^{\star T}\boldsymbol{D} = \boldsymbol{D}^T\boldsymbol{U}^\star\}$, i.e., the subspace $\mathcal{S}^\perp_{\boldsymbol{U}^\star}$ is the solution space of the linear system $\boldsymbol{U}^{\star T}\boldsymbol{D} = \boldsymbol{D}^T\boldsymbol{U}^\star$. Even though $\widehat{\boldsymbol{E}}$ may not fulfill the linear system requirements, due to the fact $\widehat{\boldsymbol{E}} \in \mathcal{S}^\perp_{\boldsymbol{U}}$, we can bound

$$\|\boldsymbol{U}^{\star T}\widehat{\boldsymbol{E}} - \widehat{\boldsymbol{E}}^T\boldsymbol{U}^\star\|_F = \|\overline{\boldsymbol{U}}^T\widehat{\boldsymbol{E}} - \widehat{\boldsymbol{E}}^T\overline{\boldsymbol{U}}\|_F \leq \frac{2\tilde{c}_0}{\lambda}\|\boldsymbol{\Delta}\|_F,$$

where the first equality holds because of $\widehat{\boldsymbol{E}} \in \mathcal{S}^\perp_{\boldsymbol{U}}$, i.e., $\boldsymbol{U}^T\widehat{\boldsymbol{E}} = \widehat{\boldsymbol{E}}^T\boldsymbol{U}$. According to Hoffman's error bound for linear system (Hoffman, 2003; Pena et al., 2021; Güler et al., 1995), there exists some constant $\kappa > 0$, which only depends on the linear system of $\boldsymbol{U}^\star$, and some $\boldsymbol{D} \in \mathcal{S}^\perp_{\boldsymbol{U}^\star}$ such that

$$\|\boldsymbol{D} - \widehat{\boldsymbol{E}}\|_F \leq \kappa\|\boldsymbol{U}^{\star T}\widehat{\boldsymbol{E}} - \widehat{\boldsymbol{E}}^T\boldsymbol{U}^\star\|_F \leq \frac{2\kappa\tilde{c}_0}{\lambda}\|\boldsymbol{\Delta}\|_F. \tag{46}$$

Then, we have

$$\begin{aligned}\|(\boldsymbol{U}^\star \widehat{\boldsymbol{E}}^T + \widehat{\boldsymbol{E}}\boldsymbol{U}^{\star T})_\Omega\|_F \geq& \|(\boldsymbol{U}^\star\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{U}^{\star T})_\Omega\|_F - 2\|\boldsymbol{U}^\star\|\|\boldsymbol{D} - \widehat{\boldsymbol{E}}\|_F \\ \geq& \sqrt{\gamma}\|\boldsymbol{D}\|_F - 2\|\boldsymbol{U}^\star\|\|\boldsymbol{D} - \widehat{\boldsymbol{E}}\|_F \geq \sqrt{\gamma} - (\sqrt{\gamma} + 2\|\boldsymbol{U}^\star\|)\|\boldsymbol{D} - \widehat{\boldsymbol{E}}\|_F,\end{aligned}$$

where the first inequality holds because of the triangle inequality and the fact $\|\boldsymbol{U}\boldsymbol{D}\|_F \leq \|\boldsymbol{U}\|\|\boldsymbol{D}\|_F$ holds for any matrices $\boldsymbol{U}$ and $\boldsymbol{D}$, the second inequality follows from the definition of γ, and the last one uses the triangle inequality and $\|\widehat{\boldsymbol{E}}\|_F = 1$. Substituting this and (46) into (45) yields

$$\sqrt{\widehat{\gamma}} \geq \|(\boldsymbol{U}^\star \widehat{\boldsymbol{E}}^T + \widehat{\boldsymbol{E}}\boldsymbol{U}^{\star T})_\Omega\|_F - \frac{2\tilde{c}_0}{\lambda}\|\boldsymbol{\Delta}\|_F \geq \sqrt{\gamma} - \frac{2\tilde{c}_0}{\lambda}\left(1 + \kappa(\sqrt{\gamma} + 2\|\boldsymbol{U}^\star\|_2)\right)\|\boldsymbol{\Delta}\|_F.$$

This, together with (11), implies $\sqrt{\widehat{\gamma}} \geq \sqrt{\gamma}/2$, which completes the proof. □

A.2.3. Discussion on the relationship between γ and λ

Note that λ is defined to satisfy $\boldsymbol{U}_k^{\star T}\boldsymbol{U}_k^\star = \sum_{i\in\mathcal{I}_k}\boldsymbol{u}_i^\star(\boldsymbol{u}_i^\star)^T \geq \lambda$. Under the setting of Gaussian matrix and noise, we could first compute the conditional expectation

$$\mathbb{E}\left[\boldsymbol{U}_k^{\star T}\boldsymbol{U}_k^\star \mid \boldsymbol{U}_k^\star \in \mathcal{C}_k\right] = \sum_{i\in I_k}\mathbb{E}[\boldsymbol{u}_i^\star(\boldsymbol{u}_i^\star)^T \mid \boldsymbol{u}_i^\star \in \mathcal{C}_k] = |\mathcal{I}_k|\mathbb{E}[\boldsymbol{u}_i^\star(\boldsymbol{u}_i^\star)^T \mid \boldsymbol{u}_i^\star \in \mathcal{C}_k].$$

Under the noiseless setting where $\mathcal{C}_k$ is an octant, a simple computation shows that

$$\mathbb{E}\left[\boldsymbol{u}_i^\star(\boldsymbol{u}_i^\star)^T \mid \boldsymbol{u}_i^\star \in \mathcal{C}_k\right] = \left(1 - \frac{2}{\pi}\right)\boldsymbol{I}_r + \frac{2}{\pi}\boldsymbol{e}\boldsymbol{e}^T,$$

where $\boldsymbol{e} \in \mathbb{R}^r$ is the all-one vector. For the noisy setting, even though it is not easy to get a close form, we could still show that $\mathbb{E}[\boldsymbol{u}_i^\star(\boldsymbol{u}_i^\star)^T \mid \boldsymbol{u}_i^\star \in \mathcal{C}_k]$ is positive definite whose least eigenvalue is independent of n and r, under the assumption $r = O(\log n)$. Besides, note that all $\boldsymbol{u}_i^\star, i \in \mathcal{I}_k$ are still i.i.d. variables under the condition $\boldsymbol{u}_i^\star \in \mathcal{C}_k$. By applying the concentration theory, we could show that $\lambda = \Theta(|I_k|) \geq \Theta(\sqrt{n})$, where the last inequality holds because we already shows that $|\mathcal{I}_k| \geq \sqrt{n}/2$ under the assumption $r = O(\log n)$.

Regarding γ, we define the linear map $\phi : \mathbb{R}^{n\times r} \mapsto \mathbb{R}^{n\times n}$ with $\phi(\boldsymbol{D}) = (\boldsymbol{U}^\star\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{U}^{\star T})_\Omega$. The linear map ϕ can be vectorized by introducing a matrix $\boldsymbol{A} \in \mathbb{R}^{n^2\times nr}$, i.e.

$$\mathrm{vec}(\phi(\boldsymbol{D})) = \boldsymbol{A}\cdot\mathrm{vec}(\boldsymbol{D})$$

where $\boldsymbol{A}$ consist of the entries of $\boldsymbol{U}^\star$. Recall that $\gamma = \min||\phi(\boldsymbol{D})||_F$ where $\boldsymbol{D} \in S^\perp, ||\boldsymbol{D}||_F = 1$. Since we already show $S^\perp$ is the kernel space of ϕ, we know γ is the smallest nonzero singular value of $\boldsymbol{A}$. By carefully computation and applying the concentration inequality, one can also show $\gamma = O(\sqrt{n})$. Then, according to Equation (7), Theorem 3.8 holds when $\|\boldsymbol{\Delta}\|_F = O(n^{\frac{1}{4}})$, in other words, we need the noise level $\delta = O(n^{-\frac{3}{4}})$.

A.3. Proofs for Section 3.3: Verification of Assumptions under the Gaussian Distribution

A.3.1. Proof of Proposition 3.10

Proof of Proposition 3.10. Since $\boldsymbol{u}_i^\star \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ for all $i \in [n]$ and $\mathcal{C}_k$ for each $k \in [2^r]$ is an orthant, we have

$$\mathbb{P}\left(\boldsymbol{u}_i^\star \in \mathcal{C}_k\right) = \frac{1}{2^r}, \forall i \in [n], \forall k \in [2^r].$$

Note that $n_k = \sum_{i=1}^n \mathbf{1}(\boldsymbol{u}_i^\star \in \mathcal{C}_k)$ denotes the number of $\boldsymbol{u}_i^\star$ belonging to orthant $\mathcal{C}_k$. Applying the Bernstein inequality (see, e.g., (Vershynin, 2018, Theorem 2.8.4)) to n_k, we obtain

$$\mathbb{P}\left(n_k - \frac{n}{2^r} \le t\right) \le \exp\left(-\frac{\frac{1}{2}t^2}{\frac{n}{2^r} + \frac{t}{3}}\right), \forall\, t \ge 0.$$

By choosing $t = n/2^{r+1} \ge \sqrt{n}/2$ due to $r \le (\log_2 n)/2$, we have for each $k \in [2^r]$,

$$\mathbb{P}\left(n_k \le \frac{n}{2^{r+1}}\right) \le \exp\left(-\frac{\sqrt{n}}{16}\right)$$

This, together with the union bound, yields

$$\mathbb{P}\left(n_k \ge \frac{\sqrt{n}}{2}, \forall k \in [2^r]\right) \le 1 - 2^r \exp\left(-\frac{\sqrt{n}}{16}\right) \le 1 - \sqrt{n}\exp\left(-\frac{\sqrt{n}}{16}\right). \tag{47}$$

Since $\sqrt{n}/2 \gg r$ and the entries of $\boldsymbol{U}^\star$ is i.i.d. standard Gaussian random variables, Assumption 3.1 always holds.

Now, we consider the near diagonal block $\Omega_{k+1,k}$ for each $k \in [2^r - 1]$. Without loss of generality, we assume $\mathcal{C}_k = \{(x_1, \cdots, x_r) : x_i \ge 0, i \in [r]\}$ and $\mathcal{C}_{k+1} = \{(x_1, \cdots, x_r) : x_i \ge 0, i \in [r-1], x_r \le 0\}$. Therefore, for each row $\boldsymbol{u}_i^\star \in \mathcal{C}_k$ and $\boldsymbol{u}_j^\star \in \mathcal{C}_{k+1}$, $\operatorname{sgn}(\boldsymbol{u}_i^\star)$ differs with $\operatorname{sgn}(\boldsymbol{u}_j^\star)$ only at the last entry. Let $\mathcal{D}_k$ consist of all the vectors $\boldsymbol{x} \in \mathcal{C}_k$ satisfying $x_r \le \max\{x_1, \cdots, x_{r-1}\}$, i.e.,

$$\mathcal{D}_k = \left\{\boldsymbol{x} \in \mathcal{C}_k : x_r \le \max\{x_1, \cdots, x_{r-1}\}\right\}.$$

One can verify that the conditional probability $\mathbb{P}(\boldsymbol{u}_i^\star \in \mathcal{D}_k \mid \boldsymbol{u}_i^\star \in \mathcal{C}_k) \ge \frac{1}{2}$, and thus

$$\mathbb{P}(\boldsymbol{u}_i^\star \in \mathcal{D}_k) = \mathbb{P}(\boldsymbol{u}_i^\star \in \mathcal{D}_k \mid \boldsymbol{u}_i^\star \in \mathcal{C}_k) \cdot \mathbb{P}(\boldsymbol{u}_i^\star \in \mathcal{C}_k) \ge \frac{1}{2^{r+1}}.$$

By applying the concentration inequality, one can show that $|\mathcal{D}_k| \ge \sqrt{n}/4$ with high probability. Then, for each $\boldsymbol{u}_i^\star \in \mathcal{D}_k$ and any $\boldsymbol{u}_j^\star \in \mathcal{C}_{k+1}$, we have

$$\mathbb{P}\left(\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\rangle \ge 0 \mid \boldsymbol{u}_i^\star\right) \ge \frac{1}{2}.$$

Note that

$$|\Omega_{k+1,k}| = \sum_{(i,j)\in\Omega_{k+1,k}} \mathbf{1}(\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\rangle \ge 0) \ge \sum_{i:\boldsymbol{u}_i^\star \in \mathcal{D}_k} \sum_{j:\boldsymbol{u}_j^\star \in \mathcal{C}_{k+1}} \mathbf{1}(\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\rangle \ge 0)$$

Conditioned on any $\boldsymbol{u}_i^\star \in \mathcal{D}_k$, $\mathbf{1}(\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\rangle \ge 0)$ for all i, j are independent of each other. Therefore, we can apply the concentration inequality to show that

$$\sum_{j:\boldsymbol{u}_j^\star \in \mathcal{C}_{k+1}} \mathbf{1}(\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\rangle \ge 0) \ge \frac{\sqrt{n}}{4}.$$

As a result, $|\Omega_{k+1,k}| \ge n/16$ holds with high probability. □

A.3.2. Proof of Proposition 3.11

For the noisy case, we need to first introduce the ϵ-net to proceed.

Definition A.2 (Net and Covering Number)**.** Let $\mathcal{S}^{r-1} = \{\boldsymbol{u} \in \mathbb{R}^r : \|\boldsymbol{u}\| = 1\}$ be the unit sphere in $\mathbb{R}^r$ and $\epsilon > 0$ be a parameter. A subset $\mathcal{N}_\epsilon \subseteq \mathcal{S}^{r-1}$ is said to be an ϵ-net of $\mathcal{S}^{r-1}$ if for any point $\boldsymbol{u} \in \mathcal{S}^{r-1}$, there exists a point $\boldsymbol{a} \in \mathcal{N}_\epsilon$ such that $\|\boldsymbol{u} - \boldsymbol{a}\| \le \epsilon$. The cardinality of the smallest ϵ-net is called the ϵ-covering number, denoted by $N(\mathcal{S}^{r-1}, \epsilon)$.

According to Szarek (1997), we know that, for any $\epsilon > 0$,

$$N(\mathcal{S}^{r-1}, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^r. \tag{48}$$

Let $\mathcal{N}_{\frac{1}{4}}$ denote the $\frac{1}{4}$-net $\mathcal{S}^{r-1}$ with the smallest cardinality. Then, (48) implies that $K = |\mathcal{N}_{\frac{1}{4}}| \leq 9^r$. Let $\mathcal{N}_{\frac{1}{4}} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M\}$, then we define

$$\mathcal{C}_k := \left\{\boldsymbol{u} \in \mathbb{R}^r : \boldsymbol{u}^T \boldsymbol{a}_k \geq \frac{\sqrt{6}+\sqrt{2}}{4}\|\boldsymbol{u}\|\right\}, \forall k \in [K].$$

Since $\boldsymbol{a}_k \in \mathcal{N}_{\frac{1}{2}} \subseteq \mathcal{S}^{r-1}$, the inequality $\boldsymbol{u}^T\boldsymbol{a}_k \geq (\sqrt{6}+\sqrt{2})\|\boldsymbol{u}\|/4$ means the angle between $\boldsymbol{u}$ and $\boldsymbol{a}_i$ is less than $\arccos\left(\frac{\sqrt{6}+\sqrt{2}}{4}\right) = \frac{\pi}{12}$. According to the definition of ϵ-net, we can show the following lemma.

Lemma A.3. $\mathbb{R}^r = \bigcup_{k\in[K]} \mathcal{C}_k$.

Proof. For any $\boldsymbol{u} \in \mathbb{R}^r \setminus \{\boldsymbol{0}\}$ and $\boldsymbol{u}/\|\boldsymbol{u}\| \in \mathcal{S}^{r-1}$, there exists some $\boldsymbol{a}_i \in \mathcal{N}_{\frac{1}{4}}$ such that $\|\boldsymbol{a}_i - \boldsymbol{u}/\|\boldsymbol{u}\|\| \leq \frac{1}{4}$, which implies

$$\left\langle \boldsymbol{a}_i, \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} \right\rangle \geq \frac{31}{32} > \frac{\sqrt{6}+\sqrt{2}}{4}.$$

This implies $\boldsymbol{u} \in \mathcal{C}_i$. □

We can now define the collection of index set as follows:

$$\mathcal{I}_k = \{i \in [n] : \boldsymbol{u}_i^\star \in \mathcal{C}_k\}, \forall k \in [K],$$

where $\boldsymbol{u}_i^\star$ is the i-th row of $\boldsymbol{U}^\star$. Lemma A.3 implies that the collection of index sets $\{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_K\}$ satisfies condition (a) in Assumption 3.6.

Proof of Proposition 3.11. First, Lemma A.3 already implies condition (a). Next, for any $k \in [K]$ and any $(i,j) \in \mathcal{I}_k \times \mathcal{I}_k$, we know that $\boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star \in \mathcal{C}_k$ according to the definition of $\mathcal{I}_k$. Owing to the property of $\mathcal{C}_k$, we know that the angle between $\boldsymbol{u}_i^\star$ and $\boldsymbol{a}_k$ is less than $\pi/12$ and so is the angle between $\boldsymbol{u}_j^\star$ and $\boldsymbol{a}_k$. This implies that the angle between $\boldsymbol{u}_i^\star$ and $\boldsymbol{u}_j^\star$ is less than $\pi/6$. This further implies

$$\boldsymbol{u}_i^{\star T}\boldsymbol{u}_j^\star \geq \cos\left(\frac{\pi}{6}\right)\|\boldsymbol{u}_i^\star\|\|\boldsymbol{u}_j^{\star T}\| = \frac{\sqrt{3}}{2}\|\boldsymbol{u}_i^\star\|\|\boldsymbol{u}_j^{\star T}\|.$$

Then, we have $m_{ij} = \boldsymbol{u}_i^{\star T}\boldsymbol{u}_j^\star + \delta_{ij} \geq 0$ due to $\|\boldsymbol{\Delta}\|_\infty < \frac{1}{2}\min_{k\in[n]}\|\boldsymbol{u}_k^\star\|_2^2$. It implies that $(i,j) \in \Omega$ for all $(i,j) \in \mathcal{I}_k \times \mathcal{I}_k$, i.e., condition (b) holds.

Next, since each row $\boldsymbol{u}_i^\star$ is an i.i.d. Gaussian vector, we have $\boldsymbol{u}_i^\star/\|\boldsymbol{u}_i^\star\|_2$ follows the uniform distribution on the unit sphere $\mathcal{S}^{r-1}$. One can evaluate the probability that $\mathbb{P}(\boldsymbol{u}_i^\star \in \mathcal{C}_k)$ in the following way: the area of $\mathcal{S}^{r-1}$ is equal to $\frac{2\pi^{r/2}}{\Gamma(r/2)}$, while the area of $\mathcal{S}^{r-1} \cap \mathcal{C}^k$ is bigger than volume of $(r-1)$-dimensional ball with radius equal to $\sin(\pi/12)$, which is $\frac{\pi^{(r-1)/2}}{\Gamma((r+1)/2)} \cdot \sin^{r-1}(\pi/12)$. This implies

$$\mathbb{P}(\boldsymbol{u}_i^\star \in \mathcal{C}_k) \geq \frac{\frac{\pi^{(r-1)/2}}{\Gamma((r+1)/2)} \cdot \sin^{r-1}(\pi/12)}{\frac{2\pi^{r/2}}{\Gamma(r/2)}} = \frac{\Gamma(r/2)}{2\sqrt{\pi}\Gamma((r+1)/2)}\sin^{r-1}(\pi/12) \geq \frac{1}{\sqrt{\pi}r}\sin^r(\pi/12),$$

where the last inequality holds because $\Gamma((r+1)/2) = \frac{r-1}{2}\Gamma((r-1)/2) \leq \frac{r}{2}\Gamma(r)$ holds for all positive integers. Since $r \leq \log n/(4\log 3)$, we have

$$\mathbb{P}\left(\boldsymbol{u}_i^\star \in \mathcal{C}_k\right) \geq \frac{1}{\sqrt{\pi}(r-1)}\sin^{r-1}(\pi/12) \geq \frac{4\log 3}{\log n} \cdot n^{\frac{\log(\sin(\pi/12))}{4\log 3}} > \frac{4\log 3}{\log n} n^{-1/3}.$$

Similar to (47), one can easily show that, with probability at least $1 - 1/n$, we have $|\mathcal{I}_k| \geq \sqrt{n}/2$ for all $k \in [K]$. Moreover, since $\sqrt{n}/2 \gg r$ and $\boldsymbol{U}^\star$ is a matrix with i.i.d. Gaussian entries, it is easy to see that $\boldsymbol{U}_k^{\star T}\boldsymbol{U}_k^\star \succeq \lambda \boldsymbol{I}_r$ holds for some $\lambda > 0$.

Finally, for any two $\mathcal{C}_k$ and $\mathcal{C}_l$ satisfying $\mathcal{C}_k \cap \mathcal{C}_l \neq \emptyset$ and for any $\boldsymbol{u}_i^\star \in \mathcal{C}_k$ and $\boldsymbol{u}_j^\star \in \mathcal{C}_l$, let $\boldsymbol{u}_0^\star \in \mathcal{C}_k \cap \mathcal{C}_l$, then the angle between $\boldsymbol{u}_i^\star$ and $\boldsymbol{u}_0^\star$ is less than $\pi/6$ and so is the angle between $\boldsymbol{u}_j^\star$ and $\boldsymbol{u}_0^\star$, which implies the angle between $\boldsymbol{u}_i^\star$ and $\boldsymbol{u}_j^\star$ is less than $\frac{\pi}{3}$. Thus, we have

$$\boldsymbol{u}_i^{\star T}\boldsymbol{u}_j^\star \geq \cos\left(\frac{\pi}{3}\right)\|\boldsymbol{u}_i^\star\|\|\boldsymbol{u}_j^{\star T}\| = \frac{1}{2}\|\boldsymbol{u}_i^\star\|\|\boldsymbol{u}_j^{\star T}\|.$$

so $m_{ij} = \boldsymbol{u}_i^{\star T}\boldsymbol{u}_j^\star + \delta_{ij} \geq 0$ given the condition that $\|\boldsymbol{\Delta}\|_\infty < \min_{k\in[n]}\|\boldsymbol{u}_k^\star\|^2/2$. It implies that $(i,j) \in \Omega$ for all $(i,j) \in \mathcal{I}_k \times \mathcal{I}_l$. For any $k, l \in [K]$, we can find a path $k_0 \to k_1 \to \cdots \to k_s$ such that $k_0 = k$, $k_s = l$ and $\mathcal{C}_{k_t} \cap \mathcal{C}_{k_{t+1}} \neq \emptyset$. This complete the whole proof.

□

B. Literature Review

There are three categories of work we review: deterministic sampling, probabilistic sampling that depends on entry values, and general matrix completion and factorization literature with close relationship to our work.

Deterministic sampling Within deterministic sampling, there is work that must make incoherence or genericity assumptions on the underlying matrix, and work that removes almost all assumptions on the matrix but either needs to leverage other properties (i.e. PSD) or ignore problematic parts of the matrix. We will start with work that does not make assumptions on the underlying matrix.

Possibly most similar to our work is (Bishop & Yu, 2014), where the authors consider deterministic sampling of PSD matrices. They require sampling of principal submatrices, which is very similar to the assumptions we have made, since under ReLU sampling, we observe the diagonal, and under the partition/permutation of (4), we also observe the diagonal blocks. To the best of our knowledge, neither of the assumptions in our paper or in (Bishop & Yu, 2014) implies the other. More specifically, Bishop & Yu (2014) consider deterministic sampling and assume that a collection of subsets of Ω admit an ordering such that any two adjacent parts have enough overlap. However, in the ReLU sampling setting, this order is hard to construct, and it is not clear whether there exists such an ordered collection of subsets. Moreover, both the analysis and the algorithm proposed in (Bishop & Yu, 2014) heavily rely on these ordered subsets, while we only use our partition for analysis. They design an algorithm for completion based on their theoretical guarantees, whereas our work simply uses the well-known gradient descent algorithm for completion. Additionally, to the best of our knowledge, it is not possible to extend their approach to non-PSD matrices. Our approach also relies heavily on the synergy of PSD matrices and ReLU sampling, but we believe that the generalization we provide in Assumption 3.6, and further generalizations thereof, can break this dependency.

The work in (Mazumdar & Rawat, 2018) studies the ReLU recovery problem in the context of neural network parameters. They provide theoretical guarantees for estimating a low-rank matrix observed after adding a bias vector and passing through a ReLU. They formulate a likelihood based on the distribution of the bias vector, and show that the maximizer of the likelihood is close to the planted low-rank matrix. However, it is not clear how to solve their proposed likelihood. Their work was inspired in part by the related work in (Soltanolkotabi, 2017), which uses gradient descent to estimate vectors that are observed through a known linear operator and then passed through a ReLU, and provides high-probability finite sample guarantees on the accuracy of the estimates when the known linear measurement operator is random. The matrix completion problem in (Ganti et al., 2015) seeks to estimate a partially observed matrix, which is a low-rank matrix observed through an unknown entry-wise monotonic and Lipschitz nonlinearity (like ReLU). They estimate both the nonlinearity and the low-rank matrix in an alternating fashion. Their guarantees are for estimating the matrix entries after the nonlinearity, as opposed to the latent low-rank matrix entries directly.

(Foucart et al., 2020) studies the non-symmetric matrix completion problem when the observation pattern is deterministic. It introduces a methodology for deriving a weighting matrix tailored to the specific sampling pattern presented. They then provide an efficient initialization scheme by making good use of the weighting matrix for the matrix completion problem. Their recovery guarantees are with respect to this weighting matrix, so if parts of the true matrix are not recoverable, the corresponding part of the weight matrix will be zero. This allows them to avoid assumptions on the underlying matrix (other than standard assumptions on the maximum value of the matrix). Other works in this direction include (Chatterjee,

2020), which proves necessary asymptotic properties of the sampling patterns for a notion of "stable recovery" when no assumptions are made on the underlying matrix. The requirements are strong and general, and so the results are somewhat limited, in particular, their results imply that no sparse deterministic pattern can guarantee stable recovery.

For works that make assumptions on the underlying matrix, (Pimentel-Alarcón et al., 2016) assumes the matrix columns (or rows) are in general position and gives nearly necessarily assumptions on the deterministic pattern. (Király et al., 2015) also assumes genericity and makes stronger assumptions. (Liu et al., 2017; 2019) develops a different assumption related to the spark of the sensing matrix or observation pattern and shows that the planted matrix uniquely fits the entries. (Shapiro et al., 2018) develops conditions for a low-rank solution to be locally unique, i.e. the unique solution in a neighborhood around that point, and makes several interesting connections to the above literature. Finally, (Singer & Cucuringu, 2010) makes a connection to rigidity theory and provides a method to determine whether a unique completion of a given partially observed matrix and a given rank is possible.

Probabilistic entry-dependent sampling There is a line of work in matrix completion that considers arbitrary sampling distributions (Foygel et al., 2011; Shamir & Shalev-Shwartz, 2014) and seeks to bound the expected loss, where expectation is taken with respect to the sampling distribution. Therefore, distinct from our work, if an entry is observed with probability zero, their results provide no guarantees for recovery. In classical statistics literature, one models the data as a random variable and the missingness mechanism as random, represented by a binary random variable, and potentially dependent on the data. In this context a commonly used taxonomy of missing data mechanisms follows (Little & Rubin, 2019), which defines MCAR, MAR, and MNAR: (i) MCAR, missing completely at random, where the missing data variable is independent of the data, e.g. entries missing uniformly at random in classical matrix completion literature fits into this category; (ii) MAR, missing at random where the missing data variable is independent of the unobserved data when conditioned on that which is observed (data and/or covariates); and (iii) MNAR, missing not at random where the missing data variable depends on that which is unobserved, even conditioned on that which is observed.

Using this probabilistic setup and terminology, (Hernández-Lobato et al., 2014) proposes a Bayesian approach to jointly infer a complete data factorization model and a missing data mechanism, also modeled with matrix factorization. They also provide approximate inference methods for the resulting intractable Bayesian inference problem. (Sportisse et al., 2020a) considers identifiability of PPCA with MNAR data and provides an estimation algorithm for the principal components in this setting. (Sportisse et al., 2020b) proposes interesting EM-style methods to estimate the joint distribution of the missing data mechanism and the data. (Ma & Zhang, 2021) identifies novel sufficient conditions such that the ground truth parameters of a parametric data model are identifiable, and maximum likelihood identifies them uniquely, in the MNAR setting.

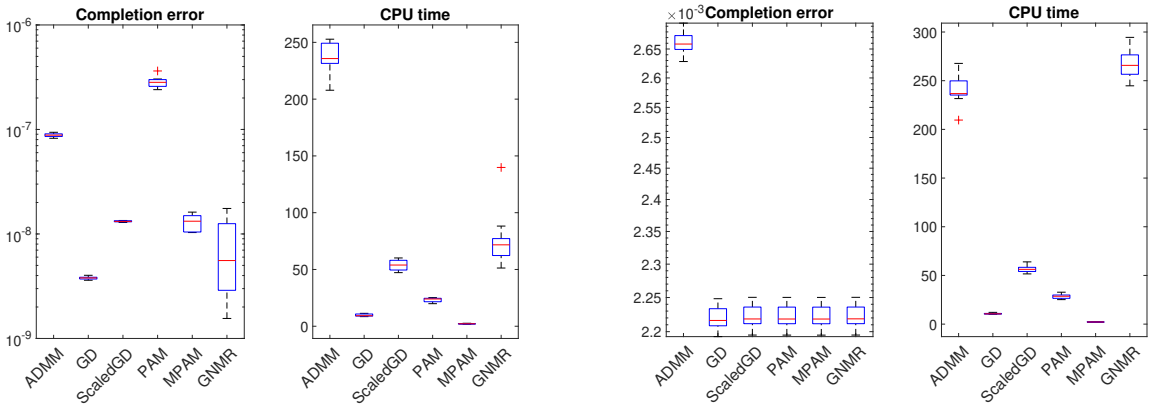
Other papers focused on the MNAR setting include (Yang et al., 2021; Jin et al., 2022; Agarwal et al., 2023). (Sengupta et al., 2023) compares matrix factorization to multiple imputation, the most popular framework in biostatistics and social sciences for imputing missing entries in the MAR setting. An empirical work that includes ReLU sampling with other MNAR sampling schemes is found in (Naik et al., 2022). The authors conclude that convex methods generally do not work as well as nonconvex in the dependent-sampling setting. Closely related to our work is also (Bhattacharya & Chatterjee, 2022), which assumes that each entry of a low-rank $\boldsymbol{M}^*$ is observed with some probability $f(\boldsymbol{M}^*)$, where f is applied entry-wise and is essentially Lipschitz and non-zero. The authors provide modified singular value thresholding and nuclear norm estimators for this setting and prove consistency in the high-dimensional regime, where the dimensions of $\boldsymbol{M}^*$ grow but the rank remains uniformly bounded. A similar estimator is proposed in (Ma & Chen, 2019), and under a low-rank assumption on the matrix of probabilities of observation, they prove error bounds for estimating the matrix of observation probabilities as well as the underlying matrix.

More general factorization and completion literature The work in (Saul, 2022) is highly related to our work, though it is not a matrix completion problem per se. This paper seeks a non-negative low-rank matrix $\mathbf{X}$ such that very sparse observation $\boldsymbol{M} = f(\mathbf{X})$ and f is a given nonlinearity, such as ReLU. They provide an EM algorithm for learning $\mathbf{X}$ but no theoretical guarantees for when this low-rank approximation exists. The work in (Seraghiti et al., 2023) studies the same problem and provides an alternating block coordinate descent approach.

Finally, the work in (Ongie et al., 2020) studies matrix completion with uniformly missing entries, but the underlying matrix columns are points on a low-dimensional nonlinear variety. The algorithm for completion lifts the partially observed data using a polynomial lifting, which creates a non-uniform sampling structure in the lifted space, where an entire deterministic set of entries in the lifted space must be missing if one entry is missing in the original space. The authors show that it is still possible to perform completion with this structured missing pattern under certain genericity conditions.

Table 2. Comparison of the norm of gradient, function value, and completion error returned by GD with different initialization.

$\sigma = 0$	α_T	β_T	γ_T
Our init	$(9.7 \pm 0.16) \cdot 10^{-7}$	$(3.3 \pm 0.36) \cdot 10^{-14}$	$(7.6 \pm 0.7) \cdot 10^{-11}$
RI init	$(6.8 \pm 0.3) \cdot 10^{-3}$	$(7.9 \pm 2.1) \cdot 10^{3}$	0.5 ± 0.13
RS init	0.1 ± 0.45	$(7.1 \pm 3.8) \cdot 10^{3}$	0.45 ± 0.24
$\sigma = 10^{-4}$	α_T	β_T	γ_T
Our init	$(9.7 \pm 0.14) \cdot 10^{-7}$	$(1.8 \pm 0.03) \cdot 10^{-4}$	$(4.4 \pm 0.13) \cdot 10^{-5}$
RI init	$(0.39 \pm 1.2) \cdot 10^{-5}$	$(8.4 \pm 2.1) \cdot 10^{3}$	0.51 ± 0.13
RS init	$(1.5 \pm 6.5) \cdot 10^{-3}$	$(8.4 \pm 2.0) \cdot 10^{3}$	0.51 ± 0.13
$\sigma = 10^{-2}$	α_T	β_T	γ_T
Our init	$(9.7 \pm 0.17) \cdot 10^{-7}$	1.8 ± 0.03	$(4.4 \pm 0.22) \cdot 10^{-3}$
RI init	$(8.1 \pm 0.35) \cdot 10^{-5}$	$(7.3 \pm 3.3) \cdot 10^{3}$	0.46 ± 0.2
RS init	$(6.4 \pm 0.19) \cdot 10^{-5}$	$(7.9 \pm 3.6) \cdot 10^{3}$	0.418 ± 0.21



(a) Noiseless case: $\sigma = 0$ (b) Noisy case: $\sigma = 10^{-2}$

Figure 8. **Completion error and CPU time of the tested methods for MC with ReLU sampling.**

C. Experimental Setups and Results

C.1. Additional results and figures

In this section, we provide the full version of Table 1, which includes noise level $\sigma = 10^{-2}$. See Table 2. We also provide an additional figure along the lines of Figure 6, also including noise level $\sigma = 10^{-2}$. In Figure 8, we see that all the algorithms perform very similarly in terms of completion accuracy with $\sigma = 10^{-2}$, and their computation time is similar to that for the noiseless case.

C.2. Experimental Setup for Section 4.3

In this subsection, we give details of the studied methods in Section 4.3.

ADMM for the convex formulation of MC. Noting that the noisy model in (1) is considered in this work, then we study the following convex formulation to complete the missing entries as studied in (Candes & Plan, 2010):

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n \times n}} \|\boldsymbol{X}\|_* \qquad \text{s.t. } \|\boldsymbol{X}_\Omega - \boldsymbol{M}_\Omega\|_F^2 \leq \delta. \tag{49}$$

In our implementation, we simply set $\delta = \|\boldsymbol{\Delta}\|_F^2$. It is obvious that the classic formulation in (Recht, 2011) is a special case of the above formulation when $\delta = 0$ in the noiseless case. Then, one can efficiently solve this convex problem using

ADMM. By introducing an auxiliary variable $\boldsymbol{Y}$, we first rewrite Problem (49) as

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times n}} \|\boldsymbol{Y}\|_* \qquad \text{s.t. } \boldsymbol{X}=\boldsymbol{Y},\ \|\boldsymbol{X}_\Omega-\boldsymbol{M}_\Omega\|_F^2\le\delta. \tag{50}$$

By introducing a dual variable $\boldsymbol{\Lambda}\in\mathbb{R}^{n\times n}$, one can write the augmented Lagrangian as follows:

$$\mathcal{L}(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\Lambda})=\|\boldsymbol{Y}\|_*+\delta_{\mathcal{X}}(\boldsymbol{X})-\langle\boldsymbol{\Lambda},\boldsymbol{X}-\boldsymbol{Y}\rangle+\frac{\rho}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2, \tag{51}$$

where $\mathcal{X}=\{\boldsymbol{X}\in\mathbb{R}^{n\times n}:\ \|\boldsymbol{X}_\Omega-\boldsymbol{M}_\Omega\|_F^2\le\delta\}$. In the implementation, we randomly genearte an initial point $(\boldsymbol{Y}^{(0)},\boldsymbol{\Lambda}^{(0)})$, whose entries are i.i.d. sampled from the standard normal distribution. Given the current iterate $(\boldsymbol{X}^{(t)},\boldsymbol{Y}^{(t)},\boldsymbol{\Lambda}^{(t)})$, ADMM generates the next iterate as follows:

- To update $\boldsymbol{X}$, we compute

  $$\boldsymbol{X}^{(t+1)}=\text{argmin}_{\boldsymbol{X}}\mathcal{L}(\boldsymbol{X},\boldsymbol{Y}^{(t)};\boldsymbol{\Lambda}^{(t)}).$$

  By letting $\boldsymbol{W}^{(t)}=\boldsymbol{Y}^{(t)}+\boldsymbol{\Lambda}^{(t)}/\rho$, we compute its closed-form solution as follows:

  $$\boldsymbol{X}^{(t+1)}=\boldsymbol{W}^{(t)},\ \text{if } \|\boldsymbol{W}_\Omega^{(t)}-\boldsymbol{M}_\Omega\|_F^2\le\delta,$$
  $$\boldsymbol{X}_{\Omega^C}^{(t+1)}=\boldsymbol{W}_{\Omega^C}^{(t)},\ \boldsymbol{X}_\Omega^{(t+1)}=\frac{\boldsymbol{W}_\Omega^{(t)}+t\boldsymbol{M}_\Omega}{1+t},\ \text{where } t=\frac{\|\boldsymbol{W}_\Omega^{(t)}-\boldsymbol{M}_\Omega\|_F^2}{\delta}-1,\ \text{otherwise.}$$

- To update $\boldsymbol{Y}$, we compute

  $$\boldsymbol{Y}^{(t+1)}=\text{argmin}_{\boldsymbol{Y}}\mathcal{L}(\boldsymbol{X}^{(t+1)},\boldsymbol{Y};\boldsymbol{\Lambda}^{(t)})=\text{argmin}_{\boldsymbol{Y}}\frac{1}{2}\left\|\boldsymbol{Y}-\left(\boldsymbol{X}^{(t+1)}-\boldsymbol{\Lambda}^{(t)}/\rho\right)\right\|_F^2+\frac{1}{\rho}\|\boldsymbol{Y}\|_*,$$

  This is to compute the proximal mapping of the nuclear norm that admits a closed-form solution.

- To update $\boldsymbol{\Lambda}$, we have

  $$\boldsymbol{\Lambda}^{(t+1)}=\boldsymbol{\Lambda}^{(t)}-\rho\left(\boldsymbol{X}^{(t+1)}-\boldsymbol{Y}^{(t+1)}\right) \tag{52}$$

We terminate the algorithm when $\left\|\boldsymbol{X}^{(t)}-\boldsymbol{Y}^{(t)}\right\|_F\le10^{-4}$ for some $t\ge0$.

ScaledGD for the formulation in (Tong et al., 2021). According to (Tong et al., 2021), one can solve the following formulation for MC:

$$\min_{\boldsymbol{L},\boldsymbol{R}\in\mathbb{R}^{n\times r}} H(\boldsymbol{L},\boldsymbol{R})=\frac{1}{2}\left\|(\boldsymbol{L}\boldsymbol{R}^T)_\Omega-\boldsymbol{M}_\Omega\right\|_F^2. \tag{53}$$

Here, we directly use their MATLAB codes downloaded from `https://github.com/Titan-Tong/ScaledGD` to implement ScaledGD for solving the above problem. Notably, we employ our tailor-designed initialization in Section 4.1 to initialize ScaledGD, which demonstrates great performance in our experiments. We terminate the algorithm when $\left\|\nabla H(\boldsymbol{L}^{(t)},\boldsymbol{R}^{(t)})\right\|_F\le10^{-4}$ for some $t\ge0$.

PAM for the formulation in (Saul, 2022). To complete sparse nonnegative matrices with low-dimensional structures, Saul (2022) studied the following formulation:

$$\min_{\boldsymbol{X},\boldsymbol{\Theta}\in\mathbb{R}^{n\times n}} \|\boldsymbol{X}-\boldsymbol{\Theta}\|_F^2 \qquad \text{s.t. } \text{rank}(\boldsymbol{\Theta})=r,\ \boldsymbol{X}_\Omega=\boldsymbol{M}_\Omega,\ \boldsymbol{X}_{\Omega^c}\le0. \tag{54}$$

Since the variables $\boldsymbol{X},\boldsymbol{\Theta}$ are separable in this problem, we can directly proximal alternating minimization (PAM) to solve this problem. In the implementation, we randomly genearte an initial point $(\boldsymbol{X}^{(0)},\boldsymbol{\Theta}^{(0)})$, whose entries are i.i.d. sampled from the standard normal distribution. Given the current iterate $(\boldsymbol{X}^{(t)},\boldsymbol{\Theta}^{(t)})$, PAM generates the next iterate as follows:

$$\boldsymbol{X}^{(t+1)}\in\text{argmin}_{\boldsymbol{X}}\|\boldsymbol{X}-\boldsymbol{\Theta}^{(t)}\|_F^2+\frac{\alpha}{2}\|\boldsymbol{X}-\boldsymbol{X}^{(t)}\|_F^2 \qquad \text{s.t. } \boldsymbol{X}_\Omega=\boldsymbol{M}_\Omega,\ \boldsymbol{X}_{\Omega^c}\le0,$$
$$\boldsymbol{\Theta}^{(t+1)}\in\text{argmin}_{\boldsymbol{\Theta}}\|\boldsymbol{X}^{(t+1)}-\boldsymbol{\Theta}\|_F^2+\frac{\beta}{2}\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^{(t)}\|_F^2 \qquad \text{s.t. } \text{rank}(\boldsymbol{\Theta})=r.$$

Obviously, the above subproblems both admit closed-form solutions. We terminate the algorithm when $\|\boldsymbol{X}^{(t+1)}-\boldsymbol{X}^{(t)}\|_F+\|\boldsymbol{\Theta}^{(t+1)}-\boldsymbol{\Theta}^{(t)}\|_F\le10^{-4}$ for some $t\ge0$.

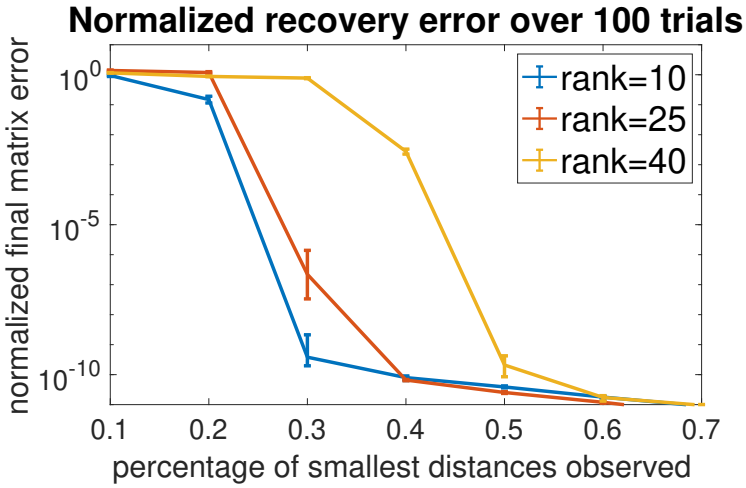


Figure 9. Matrix completion for Euclidean distance matrices where only a fraction of the smallest entries are observed. Line shows the median error, and bars show the 25% and 75% error quantiles for 100 trials.

Momentum PAM for the NMD formulation in (Seraghiti et al., 2023). Noting that $\boldsymbol{\Theta}$ in Problem (54) admits the low-rank structure, one can naturally reformulate this problem into the following nonlinear matrix decomposition (NMD) formulation:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times n},\boldsymbol{W},\boldsymbol{H}\in\mathbb{R}^{n\times r}} \|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|_F^2 \qquad \text{s.t. } \boldsymbol{X}_\Omega = \boldsymbol{M}_\Omega,\ \boldsymbol{X}_{\Omega^c} \le 0.$$

In particular, Seraghiti et al. (2023) have proposed a momentum PAM method for solving this problem. Here, we directly use their MATLAB codes downloaded from `https://gitlab.com/ngillis/ReLU-NMD` to implement momentum PAM for solving this problem. We terminate the algorithm when the relative error $\|\boldsymbol{X}^{(t)}-(\boldsymbol{W}^{(t)}\boldsymbol{H}^{(t)})_\Omega\|_F/\|\boldsymbol{X}^{(t)}\|_F \le 10^{-4}$.

Gaussian-Newton matrix recovery (GNMR) for low-rank MC (Zilber & Nadler, 2022). Zilber & Nadler (2022) employed the Gauss-Newton linearization to design a GNMR method for solving the non-convex formulation of MC. Here, we directly use the MATLAB codes in (Naik et al., 2022) to implement this method for solving MC with ReLU sampling.

C.3. Euclidean Distance Matrix Completion

Centered Euclidean distance matrices are low-rank and PSD. In applications, we may only have access to a small number of pairwise distances between points, and we would like to complete the Euclidean distance matrix to know all pairwise distances. In this section, we experiment with the scenario where we observe only the smallest distances in the matrix. We generate Euclidean distance matrices synthetically, with dimension 200, and our observation is a fraction of the smallest entries. As shown in Figure 9, depending on rank and fraction of entries, Algorithm 1 recovers the matrix exactly.

Note that after centering, 50% of entries are negative, but for a real Euclidean distance matrix none of the entries are negative. In a real data setting, we would want to design an algorithm to approximately center the matrix given only the observed entries.