Robust Noisy Inductive Matrix Completion with Local Linear Convergence

Xingcai Zhou
School of Statistics and Data Science,
and Joint Lab for Statistics and Finance
Nanjing Audit University
xczhou@nau.edu.cn

Xin Dong
School of Statistics and Data Science,
and Joint Lab for Statistics and Finance
Nanjing Audit University
MA2408005@nau.edu.cn

Linglong Kong
Department of Mathematical and Statistical Sciences
University of Alberta
lkong@ualberta.ca

Abstract

We study inductive matrix completion (IMC) in the presence of heavy-tailed and possibly asymmetric noise, where the goal is to recover a rank-r matrix of ambient dimension d given n features as side prior information. So far, there has been a lack of theoretical understanding of the statistical estimation error in noisy IMC, let alone theory under heavy-tailed noise. We develop an efficient two-stage nonconvex algorithm, called RGDIMC, via robust gradient descent with spectral initialization and feature-aware matrix factorization, based on an adaptive Huber loss to accommodate heavy-tailed noise. We prove that RGDIMC converges locally at a linear rate with sample complexity depending only linearly on n (up to $\log n$) and logarithmically on d, and achieves the minimax-optimal statistical error rate $\mathcal{O}_p(\sigma\sqrt{nr/p})$ under merely a bounded second-moment condition on the noise, where r is the rank of the unknown core matrix and p denotes the observation probability. The theoretical results are strongly supported by extensive experiments.

1 Introduction

By explicitly incorporating auxiliary features ranging from user demographics to item attributes, Inductive Matrix Completion (IMC) has become a prominent paradigm in machine learning for tackling high-dimensional relational data Natarajan and Dhillon [2014], Davenport et al. [2014], Si et al. [2016]. Unlike standard matrix recovery Keshavan et al. [2010], Recht [2011], Gross [2011] which relies solely on observed entries, IMC utilizes side information to project the target matrix onto a low-dimensional subspace, thereby enabling prediction for unseen users Xu et al. [2013], Chiang et al. [2015], Udell et al. [2016]. This capability offers a fundamental solution to the cold-start problem in recommender systems Bennett and Lanning [2007], Koren et al. [2009], where traditional methods struggle due to the lack of historical interactions.

Specifically, side information assists prediction by acting as a structural prior that bridges the gap between unseen entities and historical data. In standard matrix completion, the goal is to directly recover a highly sparse, ambient high-dimensional target matrix $\boldsymbol{L}^*$. Instead of learning isolated transductive embeddings to fill this massive matrix, IMC formulates the preference matrix as $\boldsymbol{L}^* = \boldsymbol{X}_L \boldsymbol{M}^* \boldsymbol{X}_R^\top$. This fundamentally transforms the prediction task into learning a universal, low-rank core matrix $\boldsymbol{M}^*$ that captures the latent bilinear interactions between the known feature matrices.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of theoretical guarantees.

ALGORITHM	NOISY	RATE OF ESTIMATION ERROR	SAMPLE COMPLEXITY	LINEAR CONVERGENCE
MAXIDE XU ET AL. [2013]	NO	NO	$O(rn \log n \log d)$	NO
ALTMIN JAIN AND DHILLON [2013]	NO	NO	$O(r^3 n^2 \log n \log(1/\epsilon))$	YES
GNIMCZILBER AND NADLER [2022B]	BOUNDED	NO	$O(d^2 \log n)$	NO
GDIMC ZHANG ET AL. [2018]	NO	NO	$O(r^2 n \log n \log d)$	YES
RGDIMC (OURS)	HEAVY-TAILED	$O\left(\sigma\sqrt{rn/p}\right)$	$O(rn \log d)$	YES

Despite its conceptual elegance and empirical success, the dominant optimization paradigm for IM-Ctypically formulated as empirical risk minimization with the least-squares lossrests on an idealized assumption that the observational noise follows a Gaussian distribution. This reliance renders standard IMC methods inherently fragile, directly leading to severe performance degradation in practical scenarios corrupted by heavy-tailed noise or arbitrary outliers Klopp et al. [2017], Wang and Fan [2025], where predictions diverge significantly from the ground truth.

To combat such adversarial corruptions, a natural remedy is to adopt robust formulations utilizing the ℓ_1-norm or M-estimators. While these techniques have proven highly effective in standard matrix recovery Candès et al. [2011], extending them to the inductive setting presents unique theoretical and algorithmic challenges. Theoretically, the coupling between fixed feature matrices and the core matrix distorts the optimization landscape, making critical regularity conditions like the Restricted Isometry Property (RIP) difficult to establish for non-convex algorithms Negahban and Wainwright [2011], Zilber and Nadler [2022a], Bhojanapalli et al. [2016], Ma et al. [2018]. Computationally, existing convex robust IMC variants Xu et al. [2013] suffer from prohibitive complexity in large-scale applications. Consequently, developing a provably efficient, non-convex robust algorithm capable of both zero-shot generalization and heavy-tailed noise resilience is a pressing open problem.

While these practical challenges are pronounced, the co-existence of the inductive setting and heavy-tailed noise is prevalent in real-world applications. For instance, in recommender systems, user-item interaction matrices are intrinsically sparse, requiring matrix completion anchored by user and item features. Concurrently, user ratings frequently suffer from heavy-tailed perturbations driven by bot manipulations, uncoordinated system anomalies, and extreme user behaviors. Analogous dynamics manifest in financial technologyparticularly in credit risk modeling and portfolio uncertainty quantificationwhere high-dimensional auxiliary attributes interact with heavy-tailed, non-Gaussian noise. These scenarios underscore the urgent necessity of robust, feature-aware inductive completion.

To the best of our knowledge, there is currently no work developing robust IMC for heavy-tailed noise with local linear convergence. Its development poses challenges: **(i)** is it possible to design an efficient nonconvex algorithm to achieve the desired statistical accuracy with low sample complexity; **(ii)** is it possible to recover the ground truth matrix under only a bounded second-moment condition with local linear convergence; **(iii)** is it possible to complete the matrix with the same rate of convergence of MC under the sub-Gaussian or heavy-tailed noise.

1.1 Our Contributions

The goal of the article is to answer the aforementioned questions. **(i)** We develop an adaptive Huber loss to accommodate heavy-tailed noise, propose a two-stage efficient nonconvex algorithm (RGDIMC) via a Robust Gradient Descent with spectral initialization with feature-aware matrix factorization. **(ii)** The RGDIMC converges to the true underlying matrix at a linear rate with sample complexity only linearly depending on n (up to $\log n$) and logarithmically depending on d. **(iii)** The RGDIMC achieves the minimax-optimal statistical rate $\mathcal{O}(\sigma\sqrt{rn/p})$, which is the same order as that MC in the sub-Gaussian Ma et al. [2018] or heavy-tailed noise Wang and Fan [2025]. Please refer to Section 2 for specific symbol meanings. **(iv)** Our other significant contribution lies in the strong validation of theoretical results through extensive experiments. See Table 1 for a detailed comparison of theoretical guarantees between our method and four existing IMC algorithms. It is worth noting that our RGDIMC achieves both linear rate of convergence and sample complexity linear in the feature dimension n, and minimax-optimal statistical error under the heavy-tailed noise.

1.2 Related Work

Early works in Inductive Matrix Completion (IMC) formulated the problem via convex nuclear-norm minimization to leverage side information Xu et al. [2013]. Despite their theoretical elegance, these methods suffer from high computational costs. Consequently, recent advances have pivoted to non-convex factorization paradigms Jain and Dhillon [2013], which offer superior scalability for large-scale applications. For instance, Jain et al. Jain and Dhillon [2013] proposed an alternating minimization framework, while Zilber and Nadler Zilber and Nadler [2022b] developed a spectral method with global convergence guarantees. Other structured sparsity models, such as sparse group IMC Nazarov et al. [2018], have also been explored. More recently, Yang et al. Yang and Ma [2026] established minimax-optimal guarantees for IMC under sub-Gaussian noise, showing that sample complexity scales with feature dimensions rather than ambient matrix size. However, these methods predominantly assume Gaussian noise. While specific non-Gaussian settings like Poisson noise Cao and Xie [2015] have been studied, the problem of handling heavy-tailed noise in the inductive setting remains largely underexplored.

In the broader realm of standard matrix completion, which lacks side information, robust recovery techniques have garnered significant research attention. Prominent strategies include convex approaches utilizing the ℓ_1-norm and non-convex methods employing M-estimators Sun et al. [2020], Fan et al. [2021], such as the Huber or Tukey loss. However, generalizing these robust non-convex frameworks Bhojanapalli et al. [2016], Ma et al. [2018] to the inductive setting presents unique analytical challenges. While algorithmic philosophies share high-level similarities with robust standard completion Wang and Fan [2025], the absence of side information in standard settings preserves entry-wise independence, allowing standard decoupling techniques for $\ell_{2,\infty}$ bounds. In contrast, feature matrices $\mathbf{X}_L, \mathbf{X}_R$ in IMC act as structural projectors that tightly couple observation rows and columns. Specifically, the introduction of feature matrices deteriorates the problem conditioning and complicates the gradient dynamics, posing significant obstacles to proving linear convergence. Our work bridges this gap by establishing the first rigorous convergence analysis for robust IMC formulated via the Huber loss.

1.3 Notation

For a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{x}\|_2$ denote its ℓ_1 and ℓ_2 norms, respectively. For a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_F$ and $\|\boldsymbol{A}\|_2$ represent its Frobenius norm and spectral norm. Furthermore, we define the $2,\infty$-norm as $\|\boldsymbol{A}\|_{2,\infty} = \max_i \|A_{i,:}\|_2$, which denotes the maximum ℓ_2 norm among the rows of $\boldsymbol{A}$. Similarly, $\|\boldsymbol{A}\|_\infty = \max_{i,j} |A_{i,j}|$ denotes the maximum absolute entry of $\boldsymbol{A}$. For a set of observed indices Ω, $\mathcal{P}_\Omega(\cdot)$ is the projection operator that retains entries in Ω and sets others to zero. For the ground truth matrix $\boldsymbol{M}^*$, its singular values are denoted by $\sigma_1^* \geq \sigma_2^* \geq \cdots \geq \sigma_r^* > 0$, and $\kappa = \sigma_1^*/\sigma_r^*$ is its condition number. Denote $d = \max(d_1, d_2)$ and $n = \max(n_1, n_2)$.

2 Problem Formalization and Algorithm

We consider recovering a low-rank matrix $\boldsymbol{L}^* \in \mathbb{R}^{d_1 \times d_2}$ using side information from $\boldsymbol{X}_L \in \mathbb{R}^{d_1 \times n_1}$ and $\boldsymbol{X}_R \in \mathbb{R}^{d_2 \times n_2}$. Specifically, the target matrix L^* is assumed to reside in the column spaces of these features, admitting the bilinear factorization:

$$\boldsymbol{L}^* = \boldsymbol{X}_L \boldsymbol{M}^* \boldsymbol{X}_R^\top, \tag{1}$$

where $\boldsymbol{M}^* \in \mathbb{R}^{n_1 \times n_2}$ is the unknown low-rank core matrix with $\text{rank}(\boldsymbol{M}^*) = r \ll \min(n_1, n_2)$. To simplify the presentation and without loss of generality, we assume that the feature matrices have orthonormal columns, i.e., $\boldsymbol{X}_L^\top \boldsymbol{X}_L = \boldsymbol{I}_{n_1}$ and $\boldsymbol{X}_R^\top \boldsymbol{X}_R = \boldsymbol{I}_{n_2}$.

The observed entries $\Omega \subseteq [d_1] \times [d_2]$ are sampled via a Bernoulli model with parameter $p \in (0, 1]$. For each observed index $(i, j) \in \Omega$, the observation Y_{ij} is corrupted by additive noise:

$$Y_{ij} = L_{ij}^* + \xi_{ij} = \boldsymbol{x}_{L,i}^\top \boldsymbol{M}^* \boldsymbol{x}_{R,j} + \xi_{ij}, \tag{2}$$

where $\boldsymbol{x}_{L,i}^\top$ and $\boldsymbol{x}_{R,j}^\top$ denote the i-th row of $\boldsymbol{X}_L$ and j-th row of $\boldsymbol{X}_R$, respectively, and ξ_{ij} represents the heavy-tailed noise. We focus on the challenging scenario where the noise ξ_{ij} may exhibit heavy-

Algorithm 1 Robust Gradient Descent for IMC (RGDIMC)

Input: $\mathcal{P}_\Omega(\boldsymbol{Y}), \boldsymbol{X}_L, \boldsymbol{X}_R, r, \eta, \tau, \lambda, T$.
Phase 1: Robust Initialization
Compute $\boldsymbol{M}_0 = \frac{1}{p}\boldsymbol{X}_L^\top \mathcal{P}_\Omega(\psi_\tau(\boldsymbol{Y}))\boldsymbol{X}_R$, where $\psi_\tau(x) = \text{sign}(x)\min(|x|, \tau)$.
$[\bar{\boldsymbol{U}}, \boldsymbol{\Sigma}, \bar{\boldsymbol{V}}] = \text{SVD}_r(\boldsymbol{M}_0)$ and set $(\boldsymbol{U}_0, \boldsymbol{V}_0) = (\bar{\boldsymbol{U}}\boldsymbol{\Sigma}^{1/2}, \bar{\boldsymbol{V}}\boldsymbol{\Sigma}^{1/2})$.
Phase 2: Robust Gradient Descent
for $t = 0$ **to** $T-1$ **do**
    Update $(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}) \leftarrow (\boldsymbol{U}_t, \boldsymbol{V}_t) - \eta(\nabla_{\boldsymbol{U}} f_t, \nabla_{\boldsymbol{V}} f_t)$ based on (4).
end for
Output: $\boldsymbol{M}_T = \boldsymbol{U}_T \boldsymbol{V}_T^\top$.

tailed behavior with finite but potentially very large variance, or contain sparse outliers with large magnitudes.

To ensure robustness against outliers while maintaining local smoothness for linear convergence, we employ the Huber loss $\rho_\tau(\cdot)$:

$$\rho_\tau(x) = \begin{cases} \frac{x^2}{2}, & |x| \le \tau \\ \tau|x| - \frac{\tau^2}{2}, & |x| > \tau \end{cases}. \tag{3}$$

Building on this, we seek to estimate the low-rank factors $\boldsymbol{U}$ and $\boldsymbol{V}$ by minimizing:

$$\min_{\boldsymbol{U},\boldsymbol{V}} f_\Omega(\boldsymbol{U}, \boldsymbol{V}) := \frac{1}{2p}\sum_{(i,j)\in\Omega} \rho_\tau\left(\boldsymbol{x}_{L,i}^\top \boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{x}_{R,j} - Y_{ij}\right) + \frac{\lambda}{8}\left\|\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{V}^\top\boldsymbol{V}\right\|_F^2, \tag{4}$$

where $\boldsymbol{U} \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n_2\times r}$ are the factors of the core matrix $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{V}^\top$, and $\lambda > 0$ is a regularization parameter ensuring balanced factors. The λ-penalty serves a dual purpose: theoretically, it enforces structural bounds required for Restricted Strong Convexity (RSC); empirically, it prevents global divergence under extreme outliers, ensuring stability (see sensitivity analysis in Section 4.3).

This optimization problem (4) is tricky due to its non convexity, so we adopted a two-stage approach as follows: (i)In the first stage, rather than relying on standard spectral initializationwhich severely degrades under heavy-tailed outlierswe employ an truncation operator to filter out extreme noise and yield a reliable starting point $(\mathbf{U}_0, \mathbf{V}_0)$ within the local basin of attraction. (ii)In the second stage, equipped with explicit matrix regularization to govern factor scales, we replace the standard squared loss with the Huber loss. Its bounded influence function naturally suppresses large residual spikes, ensuring smooth optimization and provable linear convergence from $(\mathbf{U}_0, \mathbf{V}_0)$ to the ground truth $\mathbf{M}^*$. See Algorithm 1.

3 Theoretical Analysis

The theoretical guarantees of our proposed method are established under a set of standard assumptions Candès et al. [2011], Tu et al. [2016], Sun and Luo [2016] that are well-established in the matrix completion literature. Collectively, these assumptions characterize the regime in which our algorithm is guaranteed to accurately recover the target matrix L^*. Specifically, we rely on the following conditions:

Assumption 3.1. (Bernoulli sampling) Each entry (i, j) is observed independently with probability $p \in (0, 1]$, that is, the entry is missing at random with probability $1 - p$.

Assumption 3.2. (Heavy-tailed noise) The noise $\{\xi_{ij}\}$ are independent, zero-mean, and have bounded variance: $\mathbb{E}[\xi_{ij}] = 0$ and $\mathbb{E}[\xi_{ij}^2] = \sigma_{ij} \le \sigma^2$.

Assumption 3.3 (Incoherence)**.** We assume the target matrix $\boldsymbol{L}^* = \boldsymbol{X}_L\boldsymbol{M}^*\boldsymbol{X}_R^\top$ and the feature matrices $\boldsymbol{X}_L, \boldsymbol{X}_R$ satisfy the following conditions:

(i) **Incoherence of Factors:** For the ground-truth factors $\boldsymbol{U}^* = \overline{\boldsymbol{U}}^*(\boldsymbol{\Sigma}^*)^{1/2}$ and $\boldsymbol{V}^* = \overline{\boldsymbol{V}}^*(\boldsymbol{\Sigma}^*)^{1/2}$ of the core matrix $\boldsymbol{M}^*$, there exists a parameter $\mu_0 \ge 1$ such that,

$$\|\boldsymbol{X}_L\boldsymbol{U}^*\|_{2,\infty} \le \sqrt{\frac{\mu_0 r\sigma_1^*}{d_1}} \text{ and } \|\boldsymbol{X}_R\boldsymbol{V}^*\|_{2,\infty} \le \sqrt{\frac{\mu_0 r\sigma_1^*}{d_2}}.$$

(ii) **Incoherence of Features:** For the feature matrices $\boldsymbol{X}_L$ and $\boldsymbol{X}_R$, there exists a parameter $\mu_1 \geq 1$ such that,

$$\|\boldsymbol{X}_L\|_{2,\infty} \leq \sqrt{\frac{\mu_1 n_1}{d_1}} \text{ and } \|\boldsymbol{X}_R\|_{2,\infty} \leq \sqrt{\frac{\mu_1 n_2}{d_2}}.$$

Theorem 3.4 (Initialization Guarantee)**.** *Suppose that Assumptions 3.1-3.3 hold. Let $\boldsymbol{M}^*$ be the underlying core matrix with its r-th singular value denoted by σ_r^*, and define $d = \max(d_1, d_2)$. Assume that the sampling rate p satisfies:*

$$p \geq \frac{C\tau^2 \mu_1^2 n_1 n_2 \log d}{(n_1+n_2)\sigma_r^{*2} d_1 d_2}. \tag{5}$$

Then, with probability at least $1 - d^{-1}$, the initial factor $(\boldsymbol{U}_0, \boldsymbol{V}_0)$ obtained from Algorithm 1 satisfies:

$$\|\boldsymbol{U}_0 \boldsymbol{V}_0^\top - \boldsymbol{M^*}\|_2 \leq \gamma \frac{\sigma_r^*}{\sqrt{r}} + C_1 \sqrt{\frac{(n_1+n_2)\sigma^2 \log d}{p}} + C_2 \tau \sqrt{\frac{(n_1+n_2)\log d}{p}}, \tag{6}$$

where $C, C_1, C_2 > 0$ are absolute constants, $\gamma \in (0,1)$ is a small numerical constant, and $\tau = C_\tau(\|\boldsymbol{L}^*\|_\infty + \sigma\sqrt{dp})$ is the truncation parameter.

Definition 3.5 (Distance with Orthogonal Alignment)**.** For any factor pair $(\boldsymbol{U}_t, \boldsymbol{V}_t)$ and the ground truth $(\boldsymbol{U}^*, \boldsymbol{V}^*)$ of the core matrix $\boldsymbol{M}^*$, the square of distance with orthogonal alignment is defined as

$$D_t^2 = \min_{\boldsymbol{R} \in \mathcal{O}_r} \left(\|\boldsymbol{U}_t - \boldsymbol{U}^* \boldsymbol{R}\|_F^2 + \|\boldsymbol{V}_t - \boldsymbol{V}^* \boldsymbol{R}\|_F^2 \right), \tag{7}$$

where $\mathcal{O}_r \coloneqq \{\boldsymbol{R} \in \mathbb{R}^{r \times r} : \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_r\}$.

Theorem 3.6 (Local Linear Convergence)**.** *Suppose that the initial estimate $(\boldsymbol{U}_0, \boldsymbol{V}_0)$ falls within a sufficiently small neighborhood of the ground truth, as guaranteed by Theorem 3.4. If the sampling rate p satisfies*

$$p \geq \frac{C\mu_1^2 r(n_1+n_2)\log d}{d_1 d_2}, \tag{8}$$

for some sufficiently large absolute constant $C > 0$ and a sufficiently small step size $\eta > 0$, the sequence $\{(\boldsymbol{U}_t, \boldsymbol{V}_t)\}_{t \geq 0}$ generated by Algorithm 1 converges linearly toward $\boldsymbol{M}^*$ with probability at least $1 - d^{-c}$ (for some absolute constant $c > 0$), and we have

$$D_{t+1}^2 \leq \rho D_t^2 + \mathcal{C}_{stat}, \tag{9}$$

where $\rho = 1 - \frac{\eta(1-\delta_{2r})\sigma_r^{*2}}{2} + \frac{48(1+\delta_{2r})\sigma_1^{*3}\eta^2}{p} + 8\sigma_1^{*3}\lambda^2\eta^2 = 1 - \mathcal{O}(\eta\sigma_r^{*2}) \in (0,1)$ is the contraction factor for the restricted isometry constant $\delta_{2r} \in (0,1)$, and $\mathcal{C}_{stat} = \frac{C\sigma_1^*\sigma^2 r(n_1+n_2)\log d \cdot \eta^2}{p}$ represents the intrinsic statistical error floor.

Remark 3.7. The parameter $\delta_{2r} \in (0,1)$ is the Restricted Isometry Property (RIP) constant of the sampling linear operator $\boldsymbol{\mathcal{A}}$ over matrices of rank up to $2r$. By selecting a sufficiently small step size η such that

$$\eta < \frac{p(1-\delta_{2r})(\sigma_r^*)^2}{96(1+\delta_{2r})(\sigma_1^*)^3 + 16p(\sigma_1^*)^3\lambda^2},$$

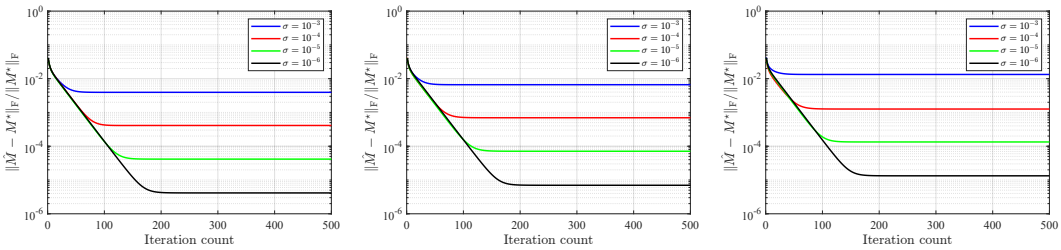
we ensure that the first-order contraction term strictly dominates the quadratic expansion terms proportional to η^2. This guarantees $\rho < 1$, thereby ensuring that the alignment error D_t^2 decays geometrically until the iteration reaches the statistical error floor $\mathcal{C}_{\text{stat}}$.

Corollary 3.8 (Minimax Optimal Estimation Error)**.** *Under the assumptions of Theorem 3.6, suppose the feature matrices $\boldsymbol{X}_L$ and $\boldsymbol{X}_R$ have orthonormal columns. If the step size is chosen as $\eta = c_\eta/\sigma_1^*$ for a sufficiently small absolute constant $c_\eta > 0$, the sequence of iterates produced by Algorithm 1 converges to a neighborhood of the true core matrix $\boldsymbol{M}^*$ with high probability. Specifically, the statistical error is bounded by*

$$\lim_{t \to \infty} \|\boldsymbol{U}_t \boldsymbol{V}_t^\top - \boldsymbol{M}^*\|_F \leq \widetilde{\mathcal{O}}\left(\sigma\sqrt{\frac{r(n_1+n_2)}{p}}\right).$$

Consequently, since the orthonormal feature matrices preserve the Frobenius norm, the recovery error for the full target matrix $\boldsymbol{L}^* = \boldsymbol{X}_L \boldsymbol{M}^* \boldsymbol{X}_R^\top$ satisfies

$$\|\hat{\boldsymbol{L}} - \boldsymbol{L}^*\|_F \leq \widetilde{\mathcal{O}}\left(\sigma\sqrt{\frac{r(n_1+n_2)}{p}}\right).$$



(a) Gaussian distribution (b) Student's distribution $t(3)$ (c) GMM

Figure 1: Relative error versus iteration count for the RGDIMC algorithm across different noise settings. Panels (a)(c) correspond to Gaussian, $t(3)$, and GMM, respectively. Curves of different colors represent varying noise levels σ.

Computational Complexity. Based on the linear convergence (Theorem 3.6), RGDIMC reaches the statistical error floor in $T = \mathcal{O}(\frac{1}{\eta\sigma_r^*}\log(\frac{D_0^2}{C_{\text{stat}}}))$ iterations. Combining the per-iteration cost of $\mathcal{O}(|\Omega|r + (d_1n_1 + d_2n_2)r)$, the total complexity is nearly linear with respect to the sample size and ambient dimensions, ensuring scalability for massive systems. See Appendix F for a detailed derivation and CPU timing comparisons.

Proof Sketch. We overcome standard Restricted Strong Convexity (RSC) failure under heavy-tailed noise via three steps. **First**, we establish RIP for the inductive operator via the Matrix Bernstein inequality. **Second**, robust initialization lands within the basin of attraction ((Theorem 3.4). **Finally**, we establish local strong convexity and smoothness of the Huber landscape. Its ℓ_1 regime truncates large residuals , bounding gradient deviations to ensure geometric contraction (Theorem 3.6) and minimax optimality (Corollary 3.8). Detailed proofs are in Appendix B - Appendix D.

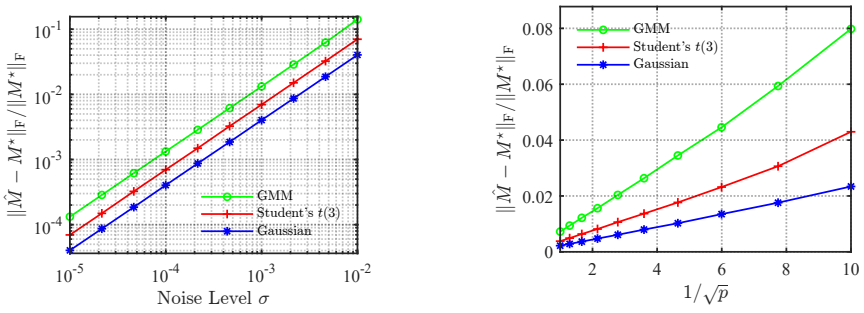
4 Experiments

In this section, we empirically evaluate the proposed RGDIMC algorithm on both synthetic and real-world datasets (MovieLens-1M and image inpainting). The primary goal is twofold: to verify our theoretical guaranteesspecifically the local linear convergence and the minimax optimal statistical rateand to assess its practical robustness against various heavy-tailed noise distributions. Comparative studies against representative standard IMC baselines (AltMin and GNIMC) demonstrate the superior efficiency and stability of our approach, especially highlighting the necessity of our robust design when standard methods fail

4.1 Simulation Experiments

In our synthetic experiments, we set the target matrix dimensions to $d_1 = d_2 = 1000$ and the ambient feature dimensions to $n_1 = n_2 = 100$. The underlying core matrix $\boldsymbol{M}^* \in \mathbb{R}^{n_1 \times n_2}$ is generated with a fixed rank of $r = 5$. Following standard practices in the optimization literature, we assume the ground-truth rank is known *a priori* to isolate and precisely evaluate the algorithmic convergence. In practical applications, the true rank can be empirically estimated via spectral methods such as Principal Component Analysis (PCA). The default observation probability is set to $p = 0.3$. To construct the core matrix $\boldsymbol{M}^* = \overline{\boldsymbol{U}}^* \boldsymbol{\Sigma}^* \overline{\boldsymbol{V}}^{*\top}$, the left and right singular vector matrices $\overline{\boldsymbol{U}}^* \in \mathbb{R}^{n_1 \times r}$ and $\overline{\boldsymbol{V}}^* \in \mathbb{R}^{n_2 \times r}$ are sampled from i.i.d. standard Gaussian distributions and subsequently orthonormalized via QR decomposition. The diagonal entries of $\boldsymbol{\Sigma}^*$, denoted by $\sigma_1^* \geq \cdots \geq \sigma_r^* > 0$, are set as linearly spaced values between r and 1. This design deliberately controls the condition number at $\kappa = \sigma_1^*/\sigma_r^* = r$, yielding a well-conditioned problem. Furthermore, the feature matrices $\boldsymbol{X}_L \in \mathbb{R}^{d_1 \times n_1}$ and $\boldsymbol{X}_R \in \mathbb{R}^{d_2 \times n_2}$ are generated similarly and orthonormalized to strictly satisfy the theoretical assumptions established in Section 3.

To evaluate the robustness of RGDIMC, we consider three distinct noise regimes for the observed entries ξ_{ij}. **(i) Standard Gaussian noise**: $\xi_{ij} \sim \mathcal{N}(0, \sigma^2)$, which is adopted as a benchmark to evaluate performance under light-tailed perturbations. **(ii) Student's t-distribution**: $\xi_{ij} \sim \sigma \cdot t(3)$,



(a) (b)

Figure 2: Statistical error analysis. (a) Relative error versus noise level σ with fixed $p = 0.3$. (b) Relative error versus $1/\sqrt{p}$ with fixed $\sigma = 10^{-3}$.

which is used to examine a heavy-tailed regime, where $t(\nu)$ is a Student's t-distribution with ν degrees of freedom. **(iii) Two-component Gaussian Mixture Model** (GMM):

$$\xi_{ij} \sim (1-\epsilon)\mathcal{N}(0, \sigma^2) + \epsilon\mathcal{N}(0, \gamma\sigma^2), \tag{10}$$

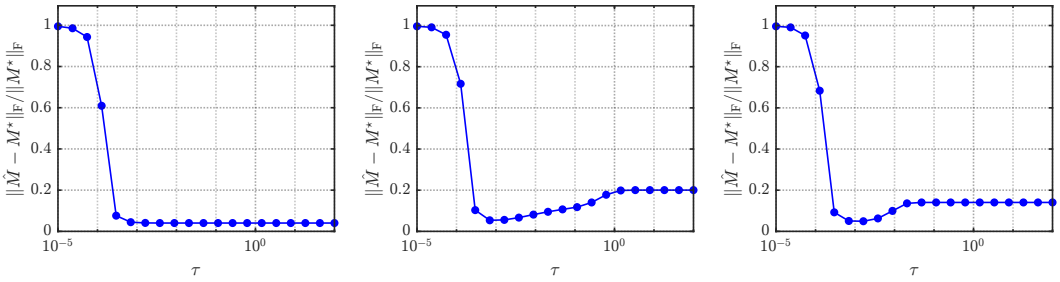
where $\epsilon \in [0, 1]$ denotes the outlier ratio and $\gamma > 1$ is a scaling factor that controls the magnitude of the outliers, which is used to simulate cases with impulsive outliers. In our simulations, fix $\epsilon = 0.1$ and $\gamma = 100$ to characterize a heavily contaminated distribution.

Regularization and Balancing. Following the established framework for non-convex robust matrix completion Wang and Fan [2025], the balancing penalty $\lambda\|\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{V}^\top\boldsymbol{V}\|_F^2$ serves a crucial structural purpose by eliminating the scaling ambiguity between factor matrices and ensuring a well-conditioned optimization landscape. Theoretically, λ does not need to adaptively scale with the noise variance or ambient dimensions; it is treated as a strictly positive constant to decouple the factor norms without dominating the primary gradient flow derived from the Huber loss. Empirically, the algorithmic performance is highly insensitive to its exact value provided it is reasonably small. Therefore, we set the default value to $\lambda = 0.1$ based on a simple grid-search on a validation set. This value is subsequently kept fixed across all heavy-tailed noise distributions, further demonstrating the parameter robustness of our RGDIMC method.

Effect of different noise distributions. We investigate the convergence behavior and robustness of the proposed RGDIMC algorithm by monitoring the relative Frobenius norm error $\|\hat{\boldsymbol{M}} - \boldsymbol{M}^*\|_F / \|\boldsymbol{M}^*\|_F$ against the iteration count. Figure 1 illustrates these results for noise intensities σ ranging from 10^{-6} to 10^{-3} across three regimes. As shown, the algorithm exhibits a consistent linear convergence rate across all regimes before plateauing at a statistical error floor. As expected, this steady-state error scales proportionally with σ, empirically validating the theoretical error floor established in Theorem 3.6. More importantly, the convergence trajectories under heavy-tailed and contaminated distributions remain nearly identical to the Gaussian baseline, underscoring the intrinsic robustness of our Huber-based approach against non-Gaussian perturbations.

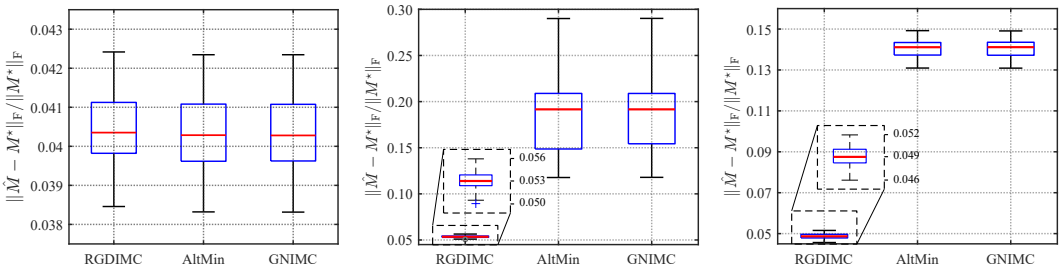
Effect of noise level σ and sampling probability p. We validate the statistical error theory of our RGDIMC. To ensure statistical reliability, all reported results are averaged over 50 independent trials, effectively accounting for the high variance inherent in heavy-tailed distributions. As shown in Figure 2, the estimation error scales linearly with noise σ and with $1/\sqrt{p}$, which is consistent with Theorem 3.6. Although the asymptotic errors are slightly higher under heavy-tailed and contaminated distributions, the consistent scaling pattern observed across all cases still highlights the inherent robustness of RGDIMC against non-Gaussian perturbations.

Sensitivity analysis of threshold τ. Figure 3 illustrates the impact of the Huber threshold τ on recovery performance. In the low-τ regime ($\tau < 10^{-4}$), excessive truncation of informative signals induces significant estimation bias, leading to high relative errors across all scenarios. As τ increases, the behavior diverges: for Gaussian noise, the error stabilizes for all $\tau > 10^{-3}$ as the loss asymptotically approaches the statistically optimal squared loss. In contrast, for heavy-tailed and contaminated noise, the curves exhibit a 'U-shape'. This occurs because an excessively large



(a) Gaussian distribution (b) Student's distribution $t(3)$ (c) Distribution in Equation (10)

Figure 3: Sensitivity analysis of threshold τ.



(a) Gaussian distribution (b) Student's distribution $t(3)$ (c) Distribution in Equation (10)

Figure 4: Performance comparison against baselines (AltMin and GNIMC).

τ degenerates the Huber penalty into the standard ℓ_2 loss, thereby stripping the algorithm of its robustness and re-exposing it to the catastrophic impact of outliers. Notably, the algorithm exhibits a broad robust range for τ, spanning approximately two orders of magnitude, suggesting that precise tuning is not required. Leveraging this insensitivity, we strictly fix $\tau = 3$ across our synthetic experiments to demonstrate the algorithmic stability without meticulous tuning. However, in practical applications where the true matrix scale and noise variance are strictly unknown, an optimal τ can be reliably determined via K-fold cross-validation, often combined with the one-standard-error rule to further guarantee robustness, as recommended by Wang and Fan [2025].

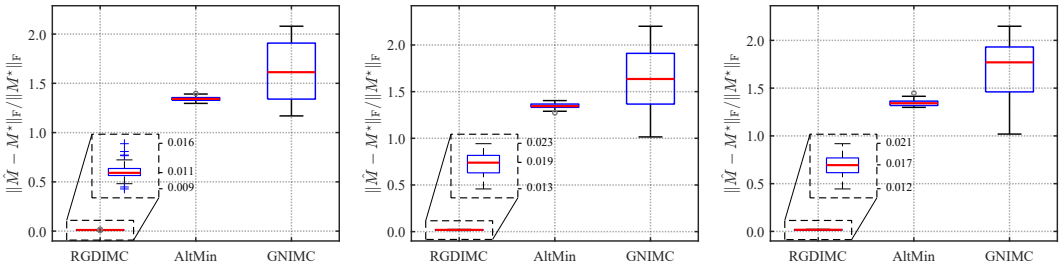
Comparison with baselines. Figure 4 compares the performance of RGDIMC against baselines (AltMin and GNIMC). Under Gaussian noise, all methods achieve comparable relative errors. However, in heavy-tailed and contaminated scenarios, RGDIMC significantly outperforms AltMin and GNIMC, which suffer from the non-robust ℓ_2 loss. In addition to its superior accuracy, our method is highly computationally efficient; a detailed comparison of execution times is provided in Appendix F.

Generalization. Figure 5 evaluates the generalization of RGDIMC. While AltMin fails due to its transductive nature and GNIMC suffers from squared-loss sensitivity to heavy-tailed noise, RGDIMC achieves near-perfect recovery. This success is rooted in the Huber loss, which prevents outliers from biasing the gradient during estimation of the weight matrix. By maintaining a stable feature-to-latent mapping, RGDIMC ensures robust extrapolation in cold-start scenarios where non-robust methods diverge.

4.2 MovieLens-1M Analysis

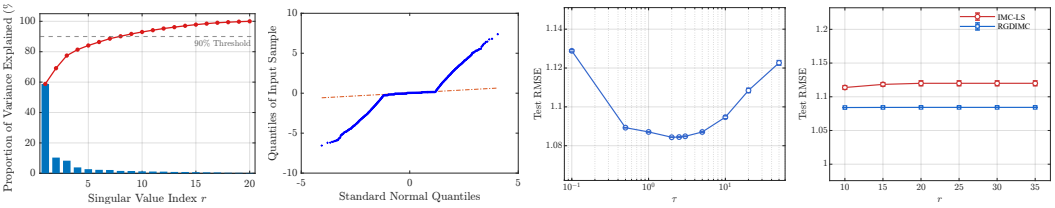
In this section, we evaluate the performance of RGDIMC on the MovieLens-1M dataset Harper and Konstan [2015] to demonstrate its efficacy in a large-scale real-world scenario. As a standard benchmark for Inductive Matrix Completion, MovieLens-1M provides rich side information for both users and movies, closely aligning with the structural assumptions of our proposed model.

The MovieLens-1M dataset consists of approximately one million ratings with scores ranging from 1 to 5, which are provided by $d_1 = 6,040$ users for $d_2 = 3,952$ movies. To construct the side infor-



(a) Gaussian distribution (b) Student's distribution $t(3)$ (c) Distribution in Equation (10)

Figure 5: Generalization to unseen users. Relative recovery errors for out-of-matrix prediction



(a) (b) (c) (d)

Figure 6: Empirical validation and sensitivity analysis on MovieLens-1M. (a) Scree plot confirming low-rank structure. (b) QQ plot of residuals revealing heavy-tailed characteristics. (c) Test RMSE versus threshold τ, reaching optimum at $\tau \approx 3$. (d) Performance stability across different ranks r compared to IMC-LS.

mation matrices $\boldsymbol{X}_L$ and $\boldsymbol{X}_R$, we encode available metadata into numerical features. Specifically, user attributes such as gender, age groups, and occupations are processed via one-hot encoding to form an $n_1 = 30$ dimensional feature space. For movies, we utilize genre information represented as multi-hot binary vectors, yielding $n_2 = 18$ dimensions.To strictly satisfy the structural prerequisites established in Section 3, we process these raw feature matrices via a two-step pipeline. First, we perform QR decomposition to extract orthonormal bases, ensuring numerical stability ($\boldsymbol{X}^\top \boldsymbol{X} = I$). Since raw sparse features yield intrinsically coherent bases, we subsequently apply a Discrete Cosine Transform (DCT). This explicitly distributes feature energy evenly, satisfying the theoretical incoherence assumption for $\boldsymbol{X}_L$ and $\boldsymbol{X}_R$ without violating orthonormality.

Figure 6a shows a sharp singular value decay, confirming the approximate low-rank structure of user preferences. Furthermore, Figure 6b reveals that the empirical residuals deviate significantly from a normal distribution, particularly at the tails, fully justifying the necessity of the robust Huber loss.To quantitatively evaluate our method, we introduce Test RMSE to measure the recovery accuracy on the unobserved set Ω_{test}; by comparing predictions against the clean ground truth, this metric objectively reflects the algorithm's ability to recover the true underlying matrix structure despite heavy-tailed noise.The effectiveness of this robust estimation is inherently tied to the Huber threshold τ. Figure 6c illustrates a clear bias-variance trade-off with respect to this parameter, exhibiting a distinct U-shaped error curve. In practical scenarios where the noise distribution is unknown, τ is determined via K-fold cross-validation. For the MovieLens-1M experiments, our grid search over $\tau \in [1, 10]$ identified that $\tau = 3$ consistently yields the optimal recovery performance. This aligns perfectly with the heavy-tailed nature of the injected Gaussian mixture noise, as it effectively caps the influence of large-magnitude outliers while preserving the gradient information of the majority of clean samples.To isolate the specific advantages of this tuned robust approach, we benchmark RGDIMC against IMC-LS [Jain and Dhillon, 2013], which is mathematically equivalent to our formulation as $\tau \to \infty$. As demonstrated in Figure 6d, RGDIMC consistently maintains a lower RMSE and superior stability across all tested ranks r. This sustained performance edge underscores that the optimized Huber loss effectively filters heavy-tailed residuals, whereas the least-squares baseline severely overfits to outliers as model complexity increases.

We also evaluate RGDIMC in a "cold-start" scenario where 20% of users are excluded from training. Our robust framework consistently achieves lower prediction errors for these unseen users compared to IMC-LS by capturing a more reliable mapping from side information to the latent space. Detailed results and convergence plots are provided in Appendix F.

4.3 Qualitative Visual Analysis

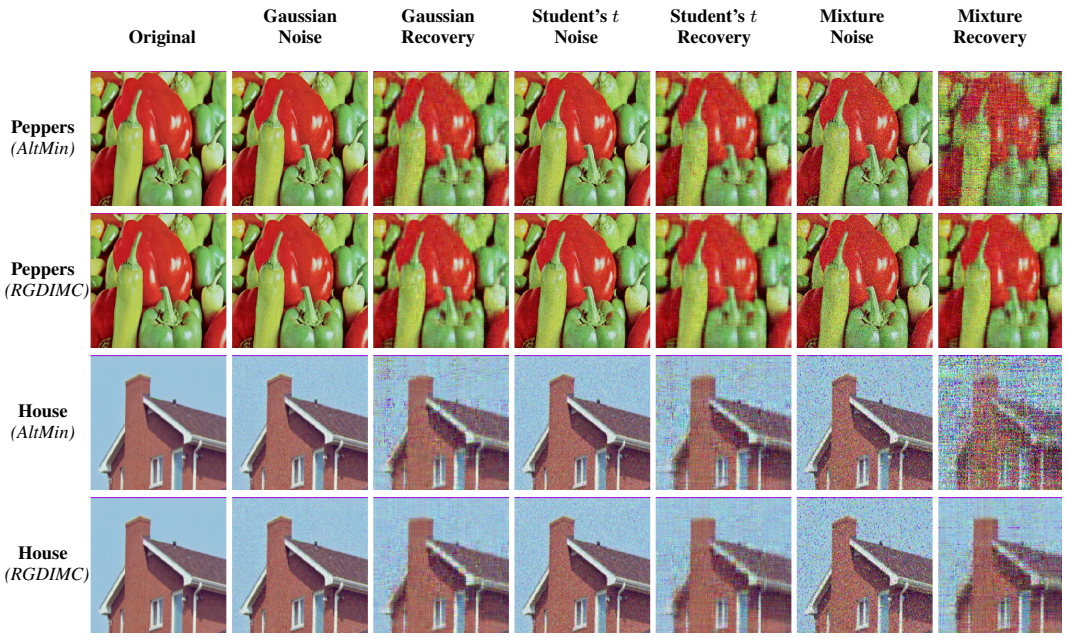


Figure 7: Visual comparison on *Peppers* and *House* images. These visual improvements are corroborated quantitatively: averaged over the three noise types, RGDIMC achieves a significantly higher PSNR of 21.98 dB on Peppers compared to 18.6 dB for AltMin, and 20.6 dB on House versus 16.7 dB for AltMin.

We further investigate the practical efficacy of RGDIMC on image inpainting using standard benchmark images (*Peppers* and *House*). As Figure 7, while standard least-squares methods such as AltMin suffer from persistent impulsive artifacts and structural blurring under heavy-tailed mixture noise, RGDIMC effectively preserves geometric details and texture homogeneity. These visual improvements are corroborated quantitatively: RGDIMC achieves significantly higher PSNR and SSIM scores compared to baselines across all noise regimes. Detailed quantitative metrics and the full visual comparison grids are deferred to Appendix Appendix F.

5 Conclusion

In this paper, we proposed RGDIMC, a robust gradient descent framework for inductive matrix completion tailored for heavy-tailed noise. By integrating the Huber loss into a non-convex optimization landscape, our method achieves resilience against impulsive outliers. We established rigorous theoretical guarantees, proving that RGDIMC converges locally at a linear rate to a statistical error floor, with merely a bounded second moment condition on the noise. RGDIMC achieves the minimax optimal statistical error rate with a sample complexity depending only linearly on the number of features. Extensive experiments on synthetic and real-world datasets, spanning recommendation systems and image inpainting, demonstrate that RGDIMC significantly outperforms standard least-squares methods under non-Gaussian perturbations while remaining competitive in Gaussian settings.

Limitations and Future Work. Our current theoretical framework focuses on stochastic heavy-tailed noise with finite variance, presenting several exciting avenues for future research. First, extending the theory to an ϵ-fraction adversarial corruption model remains a challenging open problem. Second, while our method empirically withstands infinite-variance noise due to dynamic Huber truncation, our theoretical guarantees rely on bounded second moments. Future work will aim to establish provable convergence for infinite-variance settings by leveraging fractional moments or median-of-means estimators. Algorithmically, exploring scalable Newton-type methods that exploit generalized Jacobian sparsity could yield superlinear convergence, further enhancing efficiency for large-scale systems.

Acknowledgments and Disclosure of Funding

We also sincerely thank the anonymous reviewers and the Area Chair for their constructive feedback, which significantly improved the quality of this paper. Xingcai Zhou was partially supported by the National Natural Science Foundation of China (12171242, 72342019, 12371267), the Jiangsu Provincial Key Discipline Construction Project (Statistics), and Open Project of Joint Lab for Statistics and Finance of NAU (2026JLSF204). Xin Dong was supported by the Postgraduate Research and Practice Innovation Program of Jiangsu Province (26CXJH3976). Linglong Kong was partially supported by grants from the Canada CIFAR AI Chairs program, the Alberta Machine Intelligence Institute (AMII), the Canadian AI Safety Institute Research Program at CIFAR, NSERC Alliance-Alberta Innovates Advance Program, and Natural Sciences and Engineering Council of Canada (NSERC), and Linglong Kong was also partially supported by grants from the Canada Research Chair program from NSERC.

References

James Bennett and Stan Lanning. The netflix prize. In *Proceedings of KDD Cup and Workshop*, volume 2007, page 35, 2007.

Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582. PMLR, 2016.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Yang Cao and Yao Xie. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, 2015.

Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. Matrix completion with noisy side information. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of Statistics*, 49(3):1239–1266, 2021.

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

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):1–19, 2015.

Prateek Jain and Inderjit S Dhillon. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

Olga Klopp, Karim Lounici, and Alexandre B Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1):523–564, 2017.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

Nagarajan Natarajan and Inderjit S Dhillon. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68, 2014.

Ivan Nazarov, Boris Shirokikh, Maria Burkina, Gennady Fedonin, and Maxim Panov. Sparse group inductive matrix completion. *arXiv preprint arXiv:1804.10653*, 2018.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12):3413–3430, 2011.

Si Si, Kai-Yang Chiang, Cho-Jui Hsieh, Nikhil Rao, and Inderjit S Dhillon. Goal-directed inductive matrix completion. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1165–1174, 2016.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

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

Stephen Tu, Richard Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

Bingyan Wang and Jianqing Fan. Robust matrix completion with heavy-tailed noise. *Journal of the American Statistical Association*, 120(550):922–934, 2025.

Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Yuepeng Yang and Cong Ma. Sample-efficient inductive matrix completion with noise and inexact side-information, 2026. URL `https://arxiv.org/abs/2605.17189`.

Xiao Zhang, Simon Du, and Quanquan Gu. Fast and sample efficient inductive matrix completion via multi-phase procrustes flow. In *International Conference on Machine Learning*, pages 5756–5765. PMLR, 2018.

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

Pini Zilber and Boaz Nadler. Inductive matrix completion: No bad local minima and a fast algorithm. In *International Conference on Machine Learning*, pages 27671–27692. PMLR, 2022b.

Supplementary Materials for "Robust Noisy Inductive Matrix Completion with Local Linear Convergence"

Roadmap of the Appendix. This appendix contains extended discussions, detailed mathematical proofs, and additional experimental details that support the main text. **Appendix A** provides a summary of notations, operators, and basic technical definitions used throughout the proofs. **Appendix B** introduces the core technical lemmas required for our analysis, including the Matrix Bernstein Inequality, the Restricted Isometry Property (RIP) for the inductive sampling operator, and a local norm equivalence result. **Appendix C** presents the detailed proof of the robust spectral initialization guarantee (Theorem 3.4). **Appendix D** provides the rigorous proof of the local linear convergence (Theorem 3.6) by establishing the local strong convexity and local smoothness of the optimization landscape. **Appendix E** details the derivation of the minimax optimal statistical estimation error bound (Corollary 3.8). Finally, **Appendix F** contains additional numerical experiments, including computational efficiency (CPU timing) comparisons, further quantitative visual evaluations (PSNR and SSIM) for the image inpainting tasks, an empirical sensitivity analysis of the spectral initialization to the truncation threshold τ, and a robustness analysis against inexact side information.

A Notation and Technical Definitions

Notation. Matrices and vectors are denoted by bold uppercase and lowercase letters, respectively. The inner product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimension is defined as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$, and the associated Frobenius norm is $\|\boldsymbol{A}\|_F = \sqrt{\langle \boldsymbol{A}, \boldsymbol{A} \rangle}$. Let $\Omega \subset [d_1] \times [d_2]$ be the set of observed indices. We define the orthogonal projection operator $\mathcal{P}_\Omega(\cdot)$ such that $[\mathcal{P}_\Omega(\boldsymbol{M})]_{i,j} = M_{i,j}$ if $(i,j) \in \Omega$, and 0 otherwise.

We define the linear observation operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{d_1 \times d_2}$ as $\mathcal{A}(\boldsymbol{M}) = \mathcal{P}_\Omega(\boldsymbol{X}_L \boldsymbol{M} \boldsymbol{X}_R^\top)$, where $\boldsymbol{X}_L \in \mathbb{R}^{d_1 \times n_1}$ and $\boldsymbol{X}_R \in \mathbb{R}^{d_2 \times n_2}$ are given feature matrices, and $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ is the underlying core matrix. Its adjoint operator $\mathcal{A}^* : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{n_1 \times n_2}$ is uniquely defined via the relation $\langle \mathcal{A}(\boldsymbol{M}), \boldsymbol{Y} \rangle = \langle \boldsymbol{M}, \mathcal{A}^*(\boldsymbol{Y}) \rangle$. Furthermore, we denote $\mathcal{S}_r$ as the set of all $n_1 \times n_2$ matrices with rank at most r, i.e., $\mathcal{S}_r \coloneqq \{\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2} \mid \mathrm{rank}(\boldsymbol{M}) \le r\}$.

Definition A.1 (Distance with Orthogonal Alignment)**.** For any factor pairs $(\boldsymbol{U}, \boldsymbol{V})$ and $(\boldsymbol{U}^*, \boldsymbol{V}^*)$ where $\boldsymbol{U}, \boldsymbol{U}^* \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{V}, \boldsymbol{V}^* \in \mathbb{R}^{n_2 \times r}$, the alignment error is defined as:

$$D(\boldsymbol{U}, \boldsymbol{V}; \boldsymbol{U}^*, \boldsymbol{V}^*) \coloneqq \min_{\boldsymbol{R} \in \mathcal{O}_r} \sqrt{\|\boldsymbol{U} - \boldsymbol{U}^* \boldsymbol{R}\|_F^2 + \|\boldsymbol{V} - \boldsymbol{V}^* \boldsymbol{R}\|_F^2}, \tag{A.1}$$

where $\mathcal{O}_r \coloneqq \{\boldsymbol{R} \in \mathbb{R}^{r \times r} : \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_r\}$ represents the set of $r \times r$ orthogonal matrices.

Definition A.2 (Linear Operator and RIP)**.** The observation process is modeled by the linear operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{d_1 \times d_2}$. Let $p = |\Omega|/(d_1 d_2)$ be the observation probability. The parameter $\delta_{2r} \in (0,1)$ is the Restricted Isometry Property (RIP) constant of $\mathcal{A}$, which characterizes the distortion in the Frobenius norm for any matrix $\boldsymbol{\Delta} \in \mathbb{R}^{n_1 \times n_2}$ with $\mathrm{rank}(\boldsymbol{\Delta}) \le 2r$:

$$(1 - \delta_{2r}) \|\boldsymbol{\Delta}\|_F^2 \le \frac{1}{p} \|\mathcal{A}(\boldsymbol{\Delta})\|_F^2 \le (1 + \delta_{2r}) \|\boldsymbol{\Delta}\|_F^2. \tag{A.2}$$

Definition A.3 (Aligned Parameter Error)**.** For any matrix factorization $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$, to resolve the orthogonal ambiguity, we define the aligned parameter errors as $\boldsymbol{H}_U \coloneqq \boldsymbol{U} - \boldsymbol{U}^* \boldsymbol{R}^*$ and $\boldsymbol{H}_V \coloneqq \boldsymbol{V} - \boldsymbol{V}^* \boldsymbol{R}^*$, where $\boldsymbol{R}^*$ is the optimal rotation matrix achieving the Procrustes distance:

$$\boldsymbol{R}^* = \arg \min_{\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}} \left(\|\boldsymbol{U} - \boldsymbol{U}^* \boldsymbol{R}\|_F^2 + \|\boldsymbol{V} - \boldsymbol{V}^* \boldsymbol{R}\|_F^2 \right).$$

B Technical Lemmas

Lemma B.1 (Matrix Bernstein Inequality)**.** *Consider a finite sequence $\{\boldsymbol{Z}_k\}$ of independent random matrices with dimensions $d_1 \times d_2$. Assume that each random matrix satisfies*

$$\mathbb{E}[\boldsymbol{Z}_k] = \boldsymbol{0} \quad \textit{and} \quad \|\boldsymbol{Z}_k\|_2 \le R \quad \textit{almost surely.}$$

Define the variance parameter σ^2 as

$$\sigma^2 := \max\left\{\left\|\sum_k \mathbb{E}[\boldsymbol{Z}_k \boldsymbol{Z}_k^\top]\right\|_2, \left\|\sum_k \mathbb{E}[\boldsymbol{Z}_k^\top \boldsymbol{Z}_k]\right\|_2\right\}.$$

Then, for all $t \geq 0$, the following tail bound holds:

$$\mathbb{P}\left(\left\|\sum_k \boldsymbol{Z}_k\right\|_2 \geq t\right) \leq (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

Lemma B.2 (Restricted Isometry Property for IMC)**.** *Under the incoherence conditions in Assumption 3.3, for any given RIP constant $\delta_{2r} \in (0, 1)$, there exists an absolute constant $C > 0$ such that if the sampling probability p satisfies*

$$p \geq \frac{C\mu_1^2 r(n_1 + n_2)\log d}{\delta_{2r}^2 d_1 d_2}, \tag{B.1}$$

where $d := \max(d_1, d_2)$, then with probability at least $1 - d^{-1}$, the sampling operator $\mathcal{A}$ satisfies the RIP with constant δ_{2r} for all matrices $\boldsymbol{\Delta} \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $2r$.

Proof. Our proof begins by expanding the core term $\frac{1}{p}\|\mathcal{A}(\boldsymbol{\Delta})\|_F^2$ in the RIP inequality.Using the definition of the inner product, we can rewrite it as:

$$\frac{1}{p}\|\mathcal{A}(\boldsymbol{\Delta})\|_F^2 = \langle \boldsymbol{\Delta}, \frac{1}{p}\mathcal{A}^*\mathcal{A}(\boldsymbol{\Delta})\rangle = \langle \boldsymbol{\Delta}, \left(\frac{1}{p}\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)(\boldsymbol{\Delta})\rangle + \|\boldsymbol{\Delta}\|_F^2,$$

where $\mathcal{I}$ is the identity operator. We use the Cauchy-Schwarz inequality to bound the newly emerged inner product term:

$$\langle \boldsymbol{\Delta}, \left(\frac{1}{p}\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)(\boldsymbol{\Delta})\rangle \leq \|\boldsymbol{\Delta}\|_F \left\|\left(\frac{1}{p}\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)(\boldsymbol{\Delta})\right\|_F$$

Substituting this back, we get:

$$\frac{1}{p}\|\mathcal{A}(\boldsymbol{\Delta})\|_F^2 \leq \|\boldsymbol{\Delta}\|_F \left\|\left(\frac{1}{p}\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)(\boldsymbol{\Delta})\right\|_F + \|\boldsymbol{\Delta}\|_F^2.$$

Therefore, to establish the RIP with constant δ_{2r}, it suffices to prove that the operator $\frac{1}{p}\mathcal{A}^*\mathcal{A}$ acts almost like an isometry on the restricted set of rank-$2r$ matrices. Specifically, we aim to bound the restricted operator norm:

$$\sup_{\substack{\boldsymbol{\Delta} \in \mathcal{S}_{2r} \\ \|\boldsymbol{\Delta}\|_F = 1}} \left\|\left(\frac{1}{p}\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)(\boldsymbol{\Delta})\right\|_F \leq \delta_{2r}.$$

To achieve this, we represent the operator difference as a sum of independent, zero-mean random operators. Let us define the random operator $\mathcal{Z}_{ij} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ acting on any matrix $\boldsymbol{\Delta}$ as:

$$\mathcal{Z}_{ij}(\boldsymbol{\Delta}) := \left(\frac{B_{ij}}{p} - 1\right)\langle \boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top, \boldsymbol{\Delta}\rangle \boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top$$

where B_{ij} is the Bernoulli random variable indicating whether the (i, j)-th entry is observed, and $\boldsymbol{x}_{L,i}, \boldsymbol{x}_{R,j}$ are the i-th and j-th rows of the feature matrices. It is easy to verify that $\mathbb{E}[\mathcal{Z}_{ij}] = 0$ and the entire operator difference can be expressed as:

$$\frac{1}{p}\mathcal{A}^*\mathcal{A} - \mathcal{I} = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \mathcal{Z}_{ij}.$$

Since $\mathcal{S}_{2r}$ is a non-linear manifold, we cannot apply the Matrix Bernstein inequality directly to the global operator space. Instead, following the standard covering argument, Candès et al. [2011],

we construct an ϵ-net over $\mathcal{S}_{2r}$ and bound the pointwise concentration for a fixed $\boldsymbol{\Delta} \in \mathcal{S}_{2r}$ with $\|\boldsymbol{\Delta}\|_F = 1$.

For a fixed $\boldsymbol{\Delta} \in \mathcal{S}_{2r}$ with $\|\boldsymbol{\Delta}\|_F = 1$, we evaluate the absolute maximum R and the variance parameter σ^2 of the random variables $\boldsymbol{Z}_{ij} := \mathcal{Z}_{ij}(\boldsymbol{\Delta})$. Let $\mathcal{T}$ be the tangent space at $\boldsymbol{\Delta}$. Since the restricted operator acts on $\mathcal{T}$, both the upper bound R and the variance σ^2 are fundamentally governed by the projected energy $\|\mathcal{P}_{\mathcal{T}}(\boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top)\|_F^2$.

Using the standard decomposition $\mathcal{P}_{\mathcal{T}}(\boldsymbol{M}) = \boldsymbol{P}_U\boldsymbol{M} + \boldsymbol{M}\boldsymbol{P}_V - \boldsymbol{P}_U\boldsymbol{M}\boldsymbol{P}_V$, the squared Frobenius norm of the projection is bounded by:

$$\begin{aligned}\|\mathcal{P}_{\mathcal{T}}(\boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top)\|_F^2 &\le \|\boldsymbol{P}_U\boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top\|_F^2 + \|\boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top\boldsymbol{P}_V\|_F^2 \\ &= \|\boldsymbol{P}_U\boldsymbol{x}_{L,i}\|_2^2\|\boldsymbol{x}_{R,j}\|_2^2 + \|\boldsymbol{x}_{L,i}\|_2^2\|\boldsymbol{P}_V\boldsymbol{x}_{R,j}\|_2^2.\end{aligned}$$

Under the inductive incoherence assumption, the feature rows satisfy $\|\boldsymbol{x}_{L,i}\|_2^2 \le \frac{\mu_1 n_1}{d_1}$ and $\|\boldsymbol{x}_{R,j}\|_2^2 \le \frac{\mu_1 n_2}{d_2}$. Crucially, since $\boldsymbol{P}_U$ and $\boldsymbol{P}_V$ project onto subspaces of rank at most r, the projected energy satisfies:

$$\|\boldsymbol{P}_U\boldsymbol{x}_{L,i}\|_2^2 \le \frac{\mu_1 r}{d_1}, \quad \|\boldsymbol{P}_V\boldsymbol{x}_{R,j}\|_2^2 \le \frac{\mu_1 r}{d_2}.$$

Substituting these explicit bounds into the decomposition, we obtain a unified bound on the projected maximum energy:

$$\|\mathcal{P}_{\mathcal{T}}(\boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top)\|_F^2 \le \left(\frac{\mu_1 r}{d_1}\cdot\frac{\mu_1 n_2}{d_2}\right) + \left(\frac{\mu_1 n_1}{d_1}\cdot\frac{\mu_1 r}{d_2}\right) = \frac{\mu_1^2 r(n_1+n_2)}{d_1 d_2}.$$

With this core quantity established, we can directly bound the maximum norm R. By leveraging the Cauchy-Schwarz inequality on the tangent space projection, we have:

$$\begin{aligned}R &:= \sup_{\substack{\boldsymbol{\Delta}\in\mathcal{S}_{2r}\\ \|\boldsymbol{\Delta}\|_F=1}} \|\mathcal{Z}_{ij}(\boldsymbol{\Delta})\|_F \\ &\le \frac{1}{p}\Big(\big|\langle \boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top, \boldsymbol{\Delta}\rangle\big| \cdot \|\mathcal{P}_{\mathcal{T}}(\boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top)\|_F\Big) \\ &\le \frac{1}{p}\|\mathcal{P}_{\mathcal{T}}(\boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top)\|_F^2 \\ &\le \frac{\mu_1^2 r(n_1+n_2)}{p d_1 d_2}.\end{aligned}$$

Similarly, to achieve the near-optimal sample complexity, we rigorously bound the variance parameter σ^2 defined by the maximum eigenvalue of the expected squared operator. By extracting the scalar projected energy out of the expectation, the total variance is bounded by:

$$\begin{aligned}\sigma^2 &\le \left\|\sum_{i,j}\mathbb{E}[\mathcal{Z}_{ij}^2]\right\|_2 \\ &\le \frac{1}{p}\|\mathcal{P}_{\mathcal{T}}(\boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top)\|_F^2 \cdot \sup_{\|\boldsymbol{\Delta}\|_F=1}\left\|\sum_{i,j}\langle \boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top, \boldsymbol{\Delta}\rangle \boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top\right\|_F \\ &= \frac{1}{p}\|\mathcal{P}_{\mathcal{T}}(\boldsymbol{x}_{L,i}\boldsymbol{x}_{R,j}^\top)\|_F^2 \cdot \sup_{\|\boldsymbol{\Delta}\|_F=1}\left\|(\boldsymbol{X}_L^\top\boldsymbol{X}_L)\boldsymbol{\Delta}(\boldsymbol{X}_R^\top\boldsymbol{X}_R)\right\|_F.\end{aligned}$$

By the column orthogonality of the feature matrices ($\boldsymbol{X}_L^\top\boldsymbol{X}_L = \boldsymbol{I}_{n_1}$ and $\boldsymbol{X}_R^\top\boldsymbol{X}_R = \boldsymbol{I}_{n_2}$), the operator sum perfectly collapses to $\boldsymbol{\Delta}$. Since we restrict to $\|\boldsymbol{\Delta}\|_F = 1$, its norm evaluates to exactly 1. Combined with our previously established bound for the projected energy, we arrive at the final variance bound:

$$\sigma^2 \le \frac{\mu_1^2 r(n_1+n_2)}{p d_1 d_2}.$$

With the uniform bounds for both the maximum norm R and the variance σ^2 explicitly established, we now invoke the pointwise Matrix Bernstein inequality. For a fixed $\mathbf{\Delta} \in \mathcal{S}_{2r}$ with $\|\mathbf{\Delta}\|_F = 1$, the probability that the restricted operator deviates by more than δ_{2r} is bounded by $\exp\left(-\frac{\delta_{2r}^2/2}{\sigma^2 + R\delta_{2r}/3}\right)$.

To extend this pointwise concentration to the entire non-linear manifold $\mathcal{S}_{2r}$, we employ a standard covering argument. We construct an ϵ-net over the unit sphere of $\mathcal{S}_{2r}$. The covering number (the number of points in the net) is proportional to the degrees of freedom of the rank-$2r$ manifold, which is bounded by $\exp(C'r(n_1 + n_2)\log d)$.

Taking the union bound over all points in the ϵ-net, the global probability that the restricted approximation error exceeds δ_{2r} is given by:

$$\mathbb{P}\left(\sup_{\mathbf{\Delta}\in\mathcal{S}_{2r},\|\mathbf{\Delta}\|_F=1} \left\| \left(\frac{1}{p}\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)(\mathbf{\Delta}) \right\|_F \geq \delta_{2r}\right) \leq \exp\left(C'r(n_1+n_2)\log d - \frac{\delta_{2r}^2/2}{\sigma^2 + R\delta_{2r}/3}\right).$$

For the Restricted Isometry Property to hold with high probability (e.g., failing with probability at most d^{-1}), we require the exponent to be sufficiently negative. Substituting the upper bounds of R and σ^2 into the exponent and solving for the sampling probability p, we obtain the required sample complexity condition:

$$p \geq \frac{C\mu_1^2 r(n_1+n_2)\log d}{\delta_{2r}^2 d_1 d_2}.$$

Finally, by the linearity of the operator $\left(\frac{1}{p}\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)$, the uniform bound established on the unit sphere immediately implies that:

$$\left\| \left(\frac{1}{p}\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)(\mathbf{\Delta}) \right\|_F \leq \delta_{2r}\|\mathbf{\Delta}\|_F$$

holds for all $\mathbf{\Delta} \in \mathcal{S}_{2r}$ regardless of their norm. This is analytically equivalent to the standard definition of the Restricted Isometry Property. This completes the proof of Lemma B.2. □

Lemma B.3 (Local Error Equivalence)**.** *Let $\boldsymbol{R}$ be the optimal orthogonal alignment matrix between the current iterates $(\boldsymbol{U}, \boldsymbol{V})$ and the ground truth $(\boldsymbol{U}^*, \boldsymbol{V}^*)$. Define the aligned parameter errors as $\boldsymbol{H}_U = \boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}$ and $\boldsymbol{H}_V = \boldsymbol{V} - \boldsymbol{V}^*\boldsymbol{R}$, and the observation error as $\mathbf{\Delta} = \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{U}^*(\boldsymbol{V}^*)^\top$. When the iterates of Algorithm 1 enter a local neighborhood where $\max(\|\boldsymbol{H}_U\|_2, \|\boldsymbol{H}_V\|_2) \leq \frac{1}{2}\sigma_r^*$, the squared Frobenius norm of $\mathbf{\Delta}$ is bounded by:*

$$\frac{1}{8}(\sigma_r^*)^2\left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right) \leq \|\mathbf{\Delta}\|_F^2 \leq 6(\sigma_1^*)^2\left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right).$$

Proof. First, we establish the lower bound on the error in the observation space $\|\mathbf{\Delta}\|_F^2$. Using the expansion $\mathbf{\Delta} = \boldsymbol{U}^*\boldsymbol{H}_V^\top + \boldsymbol{H}_U(\boldsymbol{V}^*)^\top + \boldsymbol{H}_U\boldsymbol{H}_V^\top$ and applying the reverse triangle inequality, we get:

$$\|\mathbf{\Delta}\|_F = \|\boldsymbol{U}^*\boldsymbol{H}_V^\top + \boldsymbol{H}_U(\boldsymbol{V}^*)^\top + \boldsymbol{H}_U\boldsymbol{H}_V^\top\|_F \geq \|\boldsymbol{U}^*\boldsymbol{H}_V^\top + \boldsymbol{H}_U(\boldsymbol{V}^*)^\top\|_F - \|\boldsymbol{H}_U\boldsymbol{H}_V^\top\|_F.$$

For the first-order term, under the optimal orthogonal alignment defined above, Tu et al. [2016] provides the following fundamental lower bound:

$$\|\boldsymbol{U}^*\boldsymbol{H}_V^\top + \boldsymbol{H}_U(\boldsymbol{V}^*)^\top\|_F \geq \frac{1}{2}\sigma_r^*(\|\boldsymbol{H}_U\|_F + \|\boldsymbol{H}_V\|_F).$$

For the second-order term, utilizing the local neighborhood assumption, we have:

$$\|\boldsymbol{H}_U\boldsymbol{H}_V^\top\|_F \leq \|\boldsymbol{H}_U\|_F\|\boldsymbol{H}_V^\top\|_2 = \|\boldsymbol{H}_U\|_F\|\boldsymbol{H}_V\|_2 \leq \frac{1}{2}\sigma_r^*\|\boldsymbol{H}_U\|_F.$$

Substituting these two bounds into the reverse triangle inequality, we obtain:

$$\|\mathbf{\Delta}\|_F \geq \frac{1}{2}\sigma_r^*(\|\boldsymbol{H}_U\|_F + \|\boldsymbol{H}_V\|_F) - \frac{1}{2}\sigma_r^*\|\boldsymbol{H}_U\|_F = \frac{1}{2}\sigma_r^*\|\boldsymbol{H}_V\|_F.$$

Squaring both sides yields $\|\boldsymbol{\Delta}\|_F^2 \geq \frac{1}{4}(\sigma_r^*)^2\|\boldsymbol{H}_V\|_F^2$. By a symmetrical argument, we also obtain $\|\boldsymbol{\Delta}\|_F^2 \geq \frac{1}{4}(\sigma_r^*)^2\|\boldsymbol{H}_U\|_F^2$. Averaging these two symmetrical bounds, we arrive at the required lower bound:

$$\|\boldsymbol{\Delta}\|_F^2 \geq \frac{1}{8}(\sigma_r^*)^2(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2).$$

Next, we establish the upper bound on $\|\boldsymbol{\Delta}\|_F^2$, which is crucial for controlling the gradient. Applying Cauchy-Schwarz inequality to the expansion of $\boldsymbol{\Delta}$, we can distribute the squared Frobenius norm:

$$\|\boldsymbol{\Delta}\|_F^2 \leq 3\left(\|\boldsymbol{U}^*\boldsymbol{H}_V^\top\|_F^2 + \|\boldsymbol{H}_U(\boldsymbol{V}^*)^\top\|_F^2 + \|\boldsymbol{H}_U\boldsymbol{H}_V^\top\|_F^2\right).$$

We bound the three terms on the right-hand side separately:

$$\begin{aligned}
\|\boldsymbol{U}^*\boldsymbol{H}_V^\top\|_F^2 &\leq \|\boldsymbol{U}^*\|_2^2\|\boldsymbol{H}_V^\top\|_F^2 = (\sigma_1^*)^2\|\boldsymbol{H}_V\|_F^2,\\
\|\boldsymbol{H}_U(\boldsymbol{V}^*)^\top\|_F^2 &\leq \|\boldsymbol{H}_U\|_F^2\|(\boldsymbol{V}^*)^\top\|_2^2 = (\sigma_1^*)^2\|\boldsymbol{H}_U\|_F^2,\\
\|\boldsymbol{H}_U\boldsymbol{H}_V^\top\|_F^2 &\leq \|\boldsymbol{H}_U\|_F^2\|\boldsymbol{H}_V^\top\|_2^2 = \|\boldsymbol{H}_U\|_F^2\|\boldsymbol{H}_V\|_2^2.
\end{aligned}$$

Substituting these bounds back and factoring out the norms, we obtain:

$$\begin{aligned}
\|\boldsymbol{\Delta}\|_F^2 &\leq 3\left((\sigma_1^*)^2\|\boldsymbol{H}_V\|_F^2 + (\sigma_1^*)^2\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_U\|_F^2\|\boldsymbol{H}_V\|_2^2\right)\\
&\leq 3\left((\sigma_1^*)^2\|\boldsymbol{H}_V\|_F^2 + \left((\sigma_1^*)^2 + \|\boldsymbol{H}_V\|_2^2\right)\|\boldsymbol{H}_U\|_F^2\right).
\end{aligned}$$

Since our neighborhood assumes $\|\boldsymbol{H}_V\|_2 \leq \frac{1}{2}\sigma_r^* \leq \sigma_1^*$, it follows that $\|\boldsymbol{H}_V\|_2^2 \leq (\sigma_1^*)^2$. Thus:

$$\|\boldsymbol{\Delta}\|_F^2 \leq 3\left((\sigma_1^*)^2\|\boldsymbol{H}_V\|_F^2 + 2(\sigma_1^*)^2\|\boldsymbol{H}_U\|_F^2\right) \leq 6(\sigma_1^*)^2\left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right).$$

This completes the proof. □

C Proof of Theorem 3.4

The core of this proof is to bound the initialization error $\left\|\mathbf{U}_0\mathbf{V}_0^\top - \boldsymbol{M}^*\right\|_2$ and arrive at the final initialization guarantee stated in Theorem 3.4. According to the truncation function $\psi_\tau(x) := \text{sign}(x)\cdot\min(|x|,\tau)$ and the triangle inequality for matrix norms, we can transform $\left\|\mathbf{U}_0\mathbf{V}_0^\top - \boldsymbol{M}^*\right\|_2$ as follows:

$$\begin{aligned}
\left\|\mathbf{U}_0\mathbf{V}_0^\top - \boldsymbol{M}^*\right\|_2 &= \left\|p^{-1}\boldsymbol{X}_L^\top\mathcal{P}_\Omega\left(\psi_\tau(\boldsymbol{L}^*+\boldsymbol{E})\right)\boldsymbol{X}_R - \boldsymbol{M}^*\right\|_2\\
&= \left\|\boldsymbol{X}_L^\top(p^{-1}\mathcal{P}_\Omega\left(\psi_\tau(\boldsymbol{L}^*+\boldsymbol{E})\right) - \boldsymbol{L}^*)\boldsymbol{X}_R\right\|_2\\
&= \left\|\boldsymbol{X}_L^\top(p^{-1}\mathcal{P}_\Omega\left(\psi_\tau\left(\boldsymbol{L}^*+\boldsymbol{E}\right)\right) - p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{L}^*\right) + p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{L}^*\right) - \boldsymbol{L}^*)\boldsymbol{X}_R\right\|_2\\
&\leq \left\|\boldsymbol{X}_L^\top(p^{-1}\mathcal{P}_\Omega\left(\psi_\tau\left(\boldsymbol{L}^*+\boldsymbol{E}\right)\right) - p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{L}^*\right))\boldsymbol{X}_R\right\|_2\\
&\quad + \left\|\boldsymbol{X}_L^\top(p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{L}^*\right) - \boldsymbol{L}^*)\boldsymbol{X}_R\right\|_2
\end{aligned}$$

For clarity, we define the two terms on the right-hand side of the above equation as the statistical error term α_1 and the noise error term α_2. Our goal is to bound these two terms separately:

$$\alpha_1 := \left\|\boldsymbol{X}_L^\top\left(p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{L}^*\right) - \boldsymbol{L}^*\right)\boldsymbol{X}_R\right\|_2$$

$$\begin{aligned}
\alpha_2 &:= \left\|\boldsymbol{X}_L^\top(p^{-1}\mathcal{P}_\Omega\left(\psi_\tau\left(\boldsymbol{L}^*+\boldsymbol{E}\right)\right) - p^{-1}\mathcal{P}_\Omega\left(\boldsymbol{L}^*\right))\boldsymbol{X}_R\right\|_2\\
&= p^{-1}\left\|\boldsymbol{X}_L^\top\mathcal{P}_\Omega\left(\psi_\tau\left(\boldsymbol{L}^*+\boldsymbol{E}\right) - \boldsymbol{L}^*\right)\boldsymbol{X}_R\right\|_2
\end{aligned}$$

According to Theorem 5.2 in Zhang et al. [2018], we obtain $\alpha_1 \leq \gamma\frac{\sigma_r^*}{\sqrt{r}}$,so our main goal is to bound the α_2.Therefor,we will turn to the more challenging noise term α_2. Based on the definition of the truncation function ψ_τ, we will split the observation set Ω into two subsets and analyze them separately.

$$\mathcal{S}_1 = \left\{(i,j)\in\Omega \mid \left|L_{i,j}^* + E_{i,j}\right| \leq \tau\right\}, \quad \mathcal{S}_2 = \Omega\setminus\mathcal{S}_1.$$

According to the definition of the truncation function ψ_τ, we can decompose the error $\psi_\tau(\boldsymbol{L}^* + \boldsymbol{E}) - \boldsymbol{L}^*$ as:

$$(\psi_\tau(\boldsymbol{L}^* + \boldsymbol{E}) - \boldsymbol{L}^*)_{ij} = \begin{cases} E_{ij}, & \text{if } (i,j) \in \mathcal{S}_1, \\ \tau \cdot \operatorname{sgn}(L^*_{ij} + E_{ij}) - L^*_{ij}, & \text{if } (i,j) \in \mathcal{S}_2. \end{cases}$$

Therefore, by the definitions of S_1 and S_2, α_2 can be written as:

$$\begin{aligned} \alpha_2 &= p^{-1} \left\| \boldsymbol{X}_L^\top (\mathcal{P}_{\mathcal{S}_1}(\boldsymbol{E}) + \mathcal{P}_{\mathcal{S}_2}(\tau \cdot \operatorname{sgn}(\boldsymbol{L}^* + \boldsymbol{E}) - \boldsymbol{L}^*)) \boldsymbol{X}_R \right\|_2 \\ &\leq p^{-1} \left\| \boldsymbol{X}_L^\top \mathcal{P}_{\mathcal{S}_1}(\boldsymbol{E}) \boldsymbol{X}_R \right\|_2 + p^{-1} \left\| \boldsymbol{X}_L^\top \mathcal{P}_{\mathcal{S}_2}(\tau \cdot \operatorname{sgn}(\boldsymbol{L}^* + \boldsymbol{E}) - \boldsymbol{L}^*) \boldsymbol{X}_R \right\|_2 \end{aligned}$$

In order to bound the norm above using the Lemma B.1, we must first calculate and bound two critical quantities: (1) the upper bound for the operator norm of each independent random matrix $\boldsymbol{Z}_{ij}$ (2) the upper bound for the total matrix variance σ^2.First, we bound $\|\boldsymbol{Z}_{ij}\|_2$ on the set S_1.

$$p^{-1} \left\| \boldsymbol{X}_L^\top \mathcal{P}_{\mathcal{S}_1}(\boldsymbol{E}) \boldsymbol{X}_R \right\|_2 = \left\| \sum_{(i,j) \in \mathcal{S}_1} p^{-1} B_{ij} E_{i,j} \boldsymbol{X}_L^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{X}_R \right\|_2 := \left\| \sum_{(i,j) \in \mathcal{S}_1} \boldsymbol{Z}_{i,j} \right\|_2$$

where $B_{ij} = 1$, if $(i,j) \in S_1$; $B_{ij} = 0$, otherwise.

On the set S_1, we have:

$$|E_{ij}| = |(L^*_{ij} + E_{ij}) - L^*_{ij}| \leq |L^*_{ij} + E_{ij}| + |L^*_{ij}| < \tau + \|\boldsymbol{L}^*\|_\infty$$

then according to the assumption of incoherence , we can get:

$$\begin{aligned} \|\boldsymbol{Z}_{ij}\|_2 &= \left\| p^{-1} B_{ij} E_{i,j} \boldsymbol{X}_L^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{X}_R \right\|_2 \\ &\leq \frac{1}{p} |E_{ij}| \cdot \left\| \boldsymbol{X}_L^\top \boldsymbol{e}_i \right\|_2 \cdot \left\| \boldsymbol{X}_R^\top \boldsymbol{e}_j \right\|_2 \\ &\leq \frac{1}{p} (\tau + \|\boldsymbol{L}^*\|_\infty) \mu_1 \sqrt{\frac{n_1 n_2}{d_1 d_2}} \end{aligned}$$

Throughout this proof, we use the truncation threshold $\tau = C_\tau(\|\boldsymbol{L}^*\|_\infty + \sigma\sqrt{dp})$ as stated in Theorem 3.4. To ensure that the true signal is preserved and only the extreme noise is truncated, the constant C_τ is chosen such that the threshold strictly upper bounds the maximum signal magnitude, i.e., $\tau \geq \|\boldsymbol{L}^*\|_\infty$. Consequently, we have $\tau + \|\boldsymbol{L}^*\|_\infty \leq 2\tau$. By absorbing the constant, the bound for the operator norm can be elegantly simplified to:

$$\|\boldsymbol{Z}_{ij}\|_2 \leq \frac{1}{p} C \tau \mu_1 \sqrt{\frac{n_1 n_2}{d_1 d_2}}$$

Then, we bound σ^2 on the set S_1 . $\boldsymbol{Z} := \sum_{(i,j) \in \mathcal{S}_1} \boldsymbol{Z}_{i,j}$.Using the column orthogonality of $\boldsymbol{X}_L$ and $\boldsymbol{X}_R$, we have the following:

$$
\begin{aligned}
\left\|\mathbb{E}\left(\boldsymbol{Z}\boldsymbol{Z}^\top\right)\right\|_2 &= \left\| \sum_{(i,j)=(1,1)}^{(d_1,d_2)} \mathbb{E}\left(\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^\top\right) \right\|_2 \\
&= \left\| \sum_{(i,j)=(1,1)}^{(d_1,d_2)} \mathbb{E}\left(p^{-2}B_{ij}^2 E_{ij}^2 \boldsymbol{X}_L^T \boldsymbol{e}_i \boldsymbol{e}_j^T \boldsymbol{X}_R \boldsymbol{X}_R^T \boldsymbol{e}_j \boldsymbol{e}_i^T \boldsymbol{X}_L\right) \right\|_2 \\
&= \left\| \sum_{(i,j)=(1,1)}^{(d_1,d_2)} p^{-2}\mathbb{E}(B_{ij}^2)\mathbb{E}(E_{ij}^2)(\boldsymbol{X}_L^T \boldsymbol{e}_i \boldsymbol{e}_j^T \boldsymbol{X}_R \boldsymbol{X}_R^T \boldsymbol{e}_j \boldsymbol{e}_i^T \boldsymbol{X}_L) \right\|_2 \\
&\leq \left\| \sum_{(i,j)=(1,1)}^{(d_1,d_2)} p^{-1}\sigma^2(\boldsymbol{X}_L^T \boldsymbol{e}_i \boldsymbol{e}_j^T \boldsymbol{X}_R \boldsymbol{X}_R^T \boldsymbol{e}_j \boldsymbol{e}_i^T \boldsymbol{X}_L) \right\|_2 \\
&\leq \left\| p^{-1}\sigma^2 \sum_{i=1}^{d_1} \left(\sum_{j=1}^{d_2} \|(\boldsymbol{X}_R)_{j,*}\|_2^2 \right) (\boldsymbol{X}_L^T \boldsymbol{e}_i \boldsymbol{e}_i^T \boldsymbol{X}_L) \right\|_2 \\
&\leq \frac{\sigma^2 n_2}{p}
\end{aligned}
$$

Similarly, we can obtain $\left\|\mathbb{E}(\boldsymbol{Z}^\top\boldsymbol{Z})\right\|_2 \leq \frac{\sigma^2 n_1}{p}$,which implies:

$$
\max\left\{\left\|\mathbb{E}(\boldsymbol{Z}^\top\boldsymbol{Z})\right\|_2, \left\|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)\right\|_2\right\} \leq \frac{n_1+n_2}{p}\sigma^2.
$$

We have now established all the conditions needed for the Matrix Bernstein inequality. Applying Matrix Bernstein (Lemma B.1) under the condition that $p \geq c\mu_1 r \log d\sigma^2/(\gamma^2 d\sigma_r^{*2})$, the error bound decomposes into a variance-driven sub-Gaussian term and a max-norm-driven sub-exponential term:

$$
\left\| \sum_{(i,j)\in\mathcal{S}_1} \boldsymbol{Z}_{ij} \right\|_2 \leq C_0 \left(\underbrace{\sqrt{\frac{2(n_1+n_2)\sigma^2\log d}{p}}}_{\text{Term 1: Variance-Driven}} + \underbrace{\frac{2C\tau\mu_1\log d}{3p}\sqrt{\frac{n_1 n_2}{d_1 d_2}}}_{\text{Term 2: Max-Norm-Driven}} \right).
$$

Here, Term 1 is driven by the cumulative variance of the noise, while Term 2 is governed by the maximum norm of a single random event. To ensure Term 1 dominates Term 2, we require:

$$
\frac{C\tau\mu_1\log d}{p}\sqrt{\frac{n_1 n_2}{d_1 d_2}} \leq \sqrt{\frac{(n_1+n_2)\sigma^2\log d}{p}}.
$$

Isolating the parameters of interest, this dominance condition is satisfied whenever the sampling rate satisfies:

$$
p \geq \frac{C\tau^2\mu_1^2 n_1 n_2 \log d}{(n_1+n_2)(\sigma_r^*)^2 d_1 d_2}.
$$

This requirement is readily implied by our sample complexity for successful initialization (Theorem 3.4). Consequently, the sub-exponential term is absorbed into the leading term, simplifying the bound on $\mathcal{S}_1$ to:

$$
\mathbb{P}\left\{ \left\| \sum_{(i,j)\in\mathcal{S}_1} \boldsymbol{Z}_{ij} \right\|_2 \geq C_1\sqrt{\frac{(n_1+n_2)\sigma^2\log d}{p}} \right\} \leq \frac{1}{d},
$$

where C_1 is a constant absorbing remaining non-dominant factors. This confirms that in the high-probability regime, the error behaves like a sum of sub-Gaussian random variables.

By the same logic, we now proceed to handle the error term on the set $\mathcal{S}_2$ using an identical method:

$$p^{-1}\left\|\boldsymbol{X}_L^\top \mathcal{P}_{\mathcal{S}_2}(\tau\cdot\operatorname{sgn}(\boldsymbol{L}^*+\boldsymbol{E})-\boldsymbol{L}^*)\boldsymbol{X}_R\right\|_2 = \left\|\sum_{(i,j)\in\mathcal{S}_2}\boldsymbol{R}_{i,j}\right\|_2,$$

where $\boldsymbol{R}_{i,j} := p^{-1}B_{ij}(\tau\cdot\operatorname{sgn}(L^*_{ij}+E_{ij})-L^*_{ij})\boldsymbol{X}_L^\top \boldsymbol{e}_i\boldsymbol{e}_j^\top \boldsymbol{X}_R$.

To apply Bernstein's inequality, we calculate the uniform bound $\|\boldsymbol{R}_{i,j}\|_2$ and the matrix variance parameter. Denoting $\boldsymbol{R}=\sum_{(i,j)\in\mathcal{S}_2}\boldsymbol{R}_{i,j}$, we have:

$$\begin{aligned}\|\boldsymbol{R}_{i,j}\|_2 &= \left\|p^{-1}B_{ij}(\tau\cdot\operatorname{sgn}(L^*_{ij}+E_{ij})-L^*_{ij})\boldsymbol{X}_L^\top \boldsymbol{e}_i\boldsymbol{e}_j^\top \boldsymbol{X}_R\right\|_2\\ &= \frac{1}{p}\left|\tau\cdot\operatorname{sgn}(L^*_{ij}+E_{ij})-L^*_{ij}\right|\cdot\left\|\boldsymbol{X}_L^\top \boldsymbol{e}_i\boldsymbol{e}_j^\top \boldsymbol{X}_R\right\|_2\\ &\le \frac{2\tau}{p}\sqrt{\frac{\mu_1^2 n_1 n_2}{d_1 d_2}},\end{aligned}$$

and for the variance parameter:

$$\begin{aligned}&\left\|\mathbb{E}\left(\boldsymbol{R}\boldsymbol{R}^\top\right)\right\|_2\\ =&\left\|\sum_{(i,j)=(1,1)}^{(d_1,d_2)}\mathbb{E}\left(\boldsymbol{R}_{ij}\boldsymbol{R}_{ij}^\top\right)\right\|_2\\ =&\left\|\sum_{(i,j)=(1,1)}^{(d_1,d_2)}\mathbb{E}\left[p^{-2}B_{ij}^2(\tau\cdot\operatorname{sgn}(L^*_{ij}+E^*_{ij})-L^*_{ij})^2(\boldsymbol{X}_L^\top \boldsymbol{e}_i\boldsymbol{e}_j^\top \boldsymbol{X}_R)(\boldsymbol{X}_R^\top \boldsymbol{e}_j\boldsymbol{e}_i^\top \boldsymbol{X}_L)\right]\right\|_2\\ \le&\left\|p^{-1}4\tau^2\sum_{i,j}\|\boldsymbol{X}_R^\top \boldsymbol{e}_j\|_2^2\cdot(\boldsymbol{X}_L^\top \boldsymbol{e}_i\boldsymbol{e}_i^\top \boldsymbol{X}_L)\right\|_2\\ \le& 4\tau^2\frac{n_2}{p}.\end{aligned}$$

Symmetrically, we bound $\|\mathbb{E}(\boldsymbol{R}^\top\boldsymbol{R})\|_2\le 4\tau^2\frac{n_1}{p}$, yielding:

$$\max\left\{\left\|\mathbb{E}(\boldsymbol{R}^\top\boldsymbol{R})\right\|_2,\left\|\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^\top)\right\|_2\right\}\le 4\tau^2\frac{n_1+n_2}{p}.$$

Substituting these parameters into Matrix Bernstein's inequality yields:

$$\mathbb{P}\left\{\left\|\sum_{(i,j)=(1,1)}^{(d_1,d_2)}\boldsymbol{R}_{ij}\right\|_2\ge C_2\tau\sqrt{\frac{(n_1+n_2)\log d}{p}}\right\}\le\frac{1}{d}.$$

Finally, combining the bound for α_1 with the bounds for the two components of α_2 on $\mathcal{S}_1$ and $\mathcal{S}_2$ via the triangle inequality, we establish the overall initialization error guarantee as stated in Theorem 3.4:

$$\left\|\mathbf{U}_0\mathbf{V}_0^\top-\boldsymbol{M}^*\right\|_2\le\gamma\frac{\sigma_r^*}{\sqrt{r}}+C_1\sqrt{\frac{(n_1+n_2)\sigma^2\log d}{p}}+C_2\tau\sqrt{\frac{(n_1+n_2)\log d}{p}}.$$

D Proof of Theorem 3.6

This section provides the rigorous proof of Theorem 3.6. The core logic is to demonstrate that, within the local neighborhood guaranteed by our initialization phase, the non-convex objective function $f_\Omega(\boldsymbol{U},\boldsymbol{V})$ exhibits favorable geometric properties. These properties are essentially inherited from the Restricted Isometry Property (RIP) of the sampling operator and the robust nature of the Huber loss. The proof proceeds according to the following logical roadmap:

(i)Local Geometry and Alignment. Since the bilinear factorization introduces rotational ambiguity, standard strong convexity does not hold. Instead, we establish a *Regularity Condition* (i.e., Restricted Strong Convexity and Smoothness up to orthogonal alignment), measuring distances via the Procrustes aligned metric defined in the previous section.**(ii)From Operator to Robust Landscape.** We prove that the RIP of the sampling operator $\mathcal{A}$, combined with the gradient clipping effect of the Huber loss, sufficiently suppresses the influence of heavy-tailed noise. This ensures that the gradient of the empirical loss points reliably towards the ground truth $\boldsymbol{M}^*$.**(iii)Linear Contraction.** By injecting the established regularity condition into the gradient descent dynamics, we demonstrate a step-wise strict contraction of the aligned parameter error $D_k^2 := \|\boldsymbol{U}^k - \boldsymbol{U}^*\boldsymbol{R}^*\|_F^2 + \|\boldsymbol{V}^k - \boldsymbol{V}^*\boldsymbol{R}^*\|_F^2$. This yields the final linear convergence rate.

D.1 Local Curvature

We begin by analyzing the local strong convexity of the total objective function $f(\boldsymbol{U}, \boldsymbol{V}) = f_{\text{huber}} + f_{\text{reg}}$. In standard smooth optimization, this is evaluated via the total Hessian quadratic form in the direction of the factor error $\boldsymbol{H} = (\boldsymbol{H}_U, \boldsymbol{H}_V)$.

Evaluating the curvature directly in the factor space is challenging due to the bilinear coupling of $\boldsymbol{U}$ and $\boldsymbol{V}$. To circumvent this, our strategy is to first establish the restricted strong convexity (RSC) in the observation space with respect to the matrix error $\boldsymbol{\Delta} = \boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{U}^*\boldsymbol{V}^{*\top}$, and subsequently map this bound back to the factor space using the geometric alignment established in Lemma B.3.

While the Huber loss is not strictly twice differentiable at the transition points $z = \pm\tau$, our theoretical analysis circumvents this ambiguity without relying on pointwise second-order differentiability. Because the first-order Huber function $\psi_\tau(z) = \rho'_\tau(z)$ is 1-Lipschitz and absolutely continuous, we establish the local curvature property strictly through the integral representation of the first-order gradient difference. By the fundamental theorem of calculus for Lebesgue integrals, the generalized derivative of $\psi_\tau(z)$ is exactly the indicator function $\mathbb{I}(|z| < \tau)$. This integral representation perfectly handles the non-differentiability at $z = \pm\tau$ since sets of measure zero do not affect the Lebesgue integral, thereby allowing us to maintain our analytical flow with absolute mathematical rigor.

Curvature of the Data-Fitting Term. We first establish the local strong convexity of the Huber loss. This property is paramount as it guarantees that the gradient descent trajectory remains within a stable basin of attraction.

Unlike the standard squared loss, the Huber loss $\rho_\tau(\cdot)$ is designed to be robust against heavy-tailed noise by thresholding large residuals. However, it is not strictly twice differentiable at the transition points $z = \pm\tau$. To resolve this theoretical ambiguity without relying on pointwise second-order differentiability, we establish the local curvature property strictly through the integral representation of the first-order gradient difference.

Let $\psi_\tau(z) = \rho'_\tau(z)$ denote the first-order Huber function. Because $\psi_\tau(z)$ is 1-Lipschitz, it is absolutely continuous. By the fundamental theorem of calculus for Lebesgue integrals, we can express the increment of the gradient along the line segment between $\boldsymbol{M}$ and $\boldsymbol{M} + \boldsymbol{\Delta}$ directly using the generalized derivative of $\psi_\tau(z)$, which is exactly the indicator function $\mathbb{I}(|z| < \tau)$. This rigorously yields:

$$\langle \nabla f_{\text{huber}}(\boldsymbol{M} + \boldsymbol{\Delta}) - \nabla f_{\text{huber}}(\boldsymbol{M}), \boldsymbol{\Delta} \rangle = \frac{1}{2p} \sum_{(i,j)\in\Omega} \int_0^1 \mathbb{I}(|z_{ij}(t)| < \tau)\, dt \cdot (\boldsymbol{x}_{L,i}^\top \boldsymbol{\Delta} \boldsymbol{x}_{R,j})^2,$$

where $z_{ij}(t) = \boldsymbol{x}_{L,i}^\top (t\boldsymbol{\Delta}) \boldsymbol{x}_{R,j} + E_{ij}$.

This integral representation is strictly well-defined, and the non-differentiability at $z = \pm\tau$ does not affect the Lebesgue integral since they constitute a set of measure zero. Under the local neighborhood guaranteed by our robust spectral initialization, the structural perturbation can be bounded by leveraging the Cauchy-Schwarz inequality and the sub-multiplicativity of the spectral norm. By explicitly substituting the incoherence bounds from Assumption 3.3, we obtain for all $t \in [0, 1]$:

$$\left|\boldsymbol{x}_{L,i}^\top (t\boldsymbol{\Delta}) \boldsymbol{x}_{R,j}\right| \le t \|\boldsymbol{x}_{L,i}\|_2 \|\boldsymbol{\Delta}\|_2 \|\boldsymbol{x}_{R,j}\|_2 \le \sqrt{\frac{\mu_1^2 n_1 n_2}{d_1 d_2}} \|\boldsymbol{\Delta}\|_2.$$

Since our robust spectral initialization explicitly guarantees that the initial estimation error $\|\boldsymbol{\Delta}\|_2$ is sufficiently small, and the truncation threshold τ scales appropriately with the signal magnitude as specified in Theorem 3.4, the product of the incoherence constants and the local radius is strictly controlled. This ensures that the structural perturbation remains bounded by half the threshold:

$$\sqrt{\frac{\mu_1^2 n_1 n_2}{d_1 d_2}}\|\boldsymbol{\Delta}\|_2 \leq \frac{\tau}{2}.$$

For any observation satisfying $|E_{ij}| \leq \frac{\tau}{2}$, the entire integration path remains inside the quadratic region of the Huber loss:

$$|z_{ij}(t)| \leq \left|\boldsymbol{x}_{L,i}^\top(t\boldsymbol{\Delta})\boldsymbol{x}_{R,j}\right| + |E_{ij}| \leq \frac{\tau}{2} + \frac{\tau}{2} = \tau.$$

Therefore, for any observation satisfying $|E_{ij}| \leq \frac{\tau}{2}$, the entire integration path remains inside the quadratic region of the Huber loss ($|z_{ij}(t)| \leq \tau$). Hence, $\mathbb{I}(|z_{ij}(t)| < \tau) = 1$ for almost every $t \in [0,1]$.

The deviation of the Huber curvature from the ideal quadratic curvature is caused only by large-noise events. Under Assumption 3.2, the noise variables have bounded variance. Chebyshev's inequality gives $\mathbb{P}\left(|E_{ij}| > \frac{\tau}{2}\right) \leq \frac{4\sigma^2}{\tau^2}$. Consequently,

$$\mathbb{P}\left(|E_{ij}| \leq \frac{\tau}{2}\right) \geq 1 - \frac{4\sigma^2}{\tau^2} =: q_\tau.$$

Taking the expectation with respect to the heavy-tailed noise distribution only, we can interchange the expectation and the finite integral by Fubini's theorem. Utilizing the fundamental property that the expectation of an indicator function equals the probability of the underlying event, we obtain:

$$\begin{aligned}\mathbb{E}\left[\int_0^1 \mathbb{I}(|z_{ij}(t)| < \tau)\, dt\right] &= \int_0^1 \mathbb{E}\left[\mathbb{I}(|z_{ij}(t)| < \tau)\right] dt \\ &= \int_0^1 \mathbb{P}\big(|z_{ij}(t)| < \tau\big) dt.\end{aligned}$$

As established above, the event $\{|E_{ij}| \leq \frac{\tau}{2}\}$ logically implies the event $\{|z_{ij}(t)| \leq \tau\}$ for all $t \in [0,1]$ under our local structural neighborhood. Therefore, by the monotonicity of probability measures, we inherently have $\mathbb{P}(|z_{ij}(t)| < \tau) \geq \mathbb{P}(|E_{ij}| \leq \frac{\tau}{2}) \geq q_\tau$. Substituting this uniform lower bound back into the integral yields:

$$\int_0^1 \mathbb{P}\big(|z_{ij}(t)| < \tau\big) dt \geq \int_0^1 q_\tau\, dt = q_\tau.$$

Conditioning on the RIP event of the sampling operator $\mathcal{A}$, it follows that:

$$\mathbb{E}\left[\langle \nabla f_{\text{huber}}(\boldsymbol{M}+\boldsymbol{\Delta}) - \nabla f_{\text{huber}}(\boldsymbol{M}), \boldsymbol{\Delta}\rangle\right] \geq \frac{q_\tau}{2p} \sum_{(i,j)\in\Omega} (\boldsymbol{x}_{L,i}^\top \boldsymbol{\Delta} \boldsymbol{x}_{R,j})^2 = \frac{q_\tau}{2p}\|\mathcal{A}(\boldsymbol{\Delta})\|_F^2.$$

Substituting the RIP lower bound $\frac{1}{p}\|\mathcal{A}(\boldsymbol{\Delta})\|_F^2 \geq (1-\delta_{2r})\|\boldsymbol{\Delta}\|_F^2$, we obtain the foundational curvature lower bound in the observation space:

$$\mathbb{E}\left[\langle \nabla f_{\text{huber}}(\boldsymbol{M}+\boldsymbol{\Delta}) - \nabla f_{\text{huber}}(\boldsymbol{M}), \boldsymbol{\Delta}\rangle\right] \geq \frac{q_\tau(1-\delta_{2r})}{2}\|\boldsymbol{\Delta}\|_F^2.$$

This bound confirms that the Huber loss maintains a strictly positive effective curvature along the error direction $\boldsymbol{\Delta}$, and the curvature reduction introduced by robust truncation is rigorously controlled by the factor q_τ.

To bridge the gap between this expected curvature and the empirical curvature, we explicitly construct a uniform concentration argument. Let $\mathcal{E}(\boldsymbol{\Delta})$ denote the empirical curvature form and

$\mathbb{E}[\mathcal{E}(\boldsymbol{\Delta})]$ be its expectation. We aim to uniformly bound the deviation over the set of rank-$2r$ matrices:

$$\sup_{\boldsymbol{\Delta}\in\mathcal{S}_{2r},\|\boldsymbol{\Delta}\|_F=1} |\mathcal{E}(\boldsymbol{\Delta}) - \mathbb{E}[\mathcal{E}(\boldsymbol{\Delta})]|$$

To rigorously bound the summands in our concentration argument, we first establish an infinity-norm bound for the projection of any normalized matrix $\boldsymbol{\Delta}$ in the local restricted set $\mathcal{S}_{2r}$.

For a fixed direction $\boldsymbol{\Delta}$ with $\|\boldsymbol{\Delta}\|_F = 1$, the empirical curvature is a sum of independent random variables $W_{ij} = \frac{1}{2p} B_{ij}\mathbb{I}(|z_{ij}(t)| < \tau)(\boldsymbol{x}_{L,i}^\top \boldsymbol{\Delta}\boldsymbol{x}_{R,j})^2$. Despite the heavy-tailed noise, the active Huber indicator function inherently bounds the effective noise, and the incoherence condition (Assumption 3.3) ensures that the structural energy is uniformly bounded:

$$\sup_{\boldsymbol{\Delta}\in\mathcal{S}_{2r},\|\boldsymbol{\Delta}\|_F=1} (\boldsymbol{x}_{L,i}^\top \boldsymbol{\Delta}\boldsymbol{x}_{R,j})^2 \le \frac{\mu_1^2 r(n_1+n_2)}{d_1 d_2}.$$

This absolute boundedness allows us to legitimately apply Hoeffding's inequality. For any fixed $\boldsymbol{\Delta}$ and a given tolerance $\epsilon_0 > 0$, the deviation is bounded by:

$$\mathbb{P}\left(|\mathcal{E}(\boldsymbol{\Delta}) - \mathbb{E}[\mathcal{E}(\boldsymbol{\Delta})]| \ge \epsilon_0\right) \le 2\exp\left(-\frac{cpd_1d_2\epsilon_0^2}{\mu_1^2 r(n_1+n_2)}\right),$$

where $c > 0$ is an absolute constant.

To extend this pointwise concentration to the entire non-linear manifold $\mathcal{S}_{2r}$, we employ an ϵ-net covering argument. Let $\mathcal{N}_{1/4}$ be a $(1/4)$-net covering the unit sphere of $\mathcal{S}_{2r}$ with respect to the Frobenius norm. A standard volumetric argument bounds its cardinality by $|\mathcal{N}_{1/4}| \le 9^{2r(n_1+n_2)}$. Taking the union bound over all points in $\mathcal{N}_{1/4}$, we have:

$$\mathbb{P}\left(\sup_{\boldsymbol{\Delta}\in\mathcal{N}_{1/4}} |\mathcal{E}(\boldsymbol{\Delta}) - \mathbb{E}[\mathcal{E}(\boldsymbol{\Delta})]| \ge \epsilon_0\right) \le 2\exp\left(2r(n_1+n_2)\log 9 - \frac{cpd_1d_2\epsilon_0^2}{\mu_1^2 r(n_1+n_2)}\right).$$

Under our sample complexity assumption $p \gtrsim \frac{\mu_1^2 r^2 (n_1+n_2)^2 \log d}{d_1 d_2 \epsilon_0^2}$, the exponent is strictly negative, meaning the event holds with probability at least $1 - d^{-c'}$.

Finally, by the standard successive approximation argument on the unit sphere, the supremum over the entire set $\mathcal{S}_{2r}$ is bounded by expanding the error from the net:

$$\sup_{\boldsymbol{\Delta}\in\mathcal{S}_{2r},\|\boldsymbol{\Delta}\|_F=1} |\mathcal{E}(\boldsymbol{\Delta}) - \mathbb{E}[\mathcal{E}(\boldsymbol{\Delta})]| \le \frac{1}{1-2(1/4)} \sup_{\boldsymbol{\Delta}\in\mathcal{N}_{1/4}} |\mathcal{E}(\boldsymbol{\Delta}) - \mathbb{E}[\mathcal{E}(\boldsymbol{\Delta})]| \le 2\epsilon_0.$$

Setting the tolerance $\epsilon_0 = \frac{q_\tau(1-\delta_{2r})}{8}$, the empirical curvature $\mathcal{E}(\boldsymbol{\Delta})$ deviates from its expectation by at most a fraction of the lower bound. Subtracting this maximal uniform deviation from the expected bound (D.1), we rigorously establish the local restricted strong convexity for the empirical Huber loss with high probability.

To transition this strong convexity to the factorized parameter space $(\boldsymbol{U}, \boldsymbol{V})$, we invoke the geometric alignment from Lemma B.3. Specifically, the Frobenius norm of the prediction error $\boldsymbol{\Delta}$ is lower-bounded by the factor errors $\boldsymbol{H}_U$ and $\boldsymbol{H}_V$:

$$\|\boldsymbol{\Delta}\|_F^2 \ge \frac{1}{8}\sigma_r^{*2}(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2).$$

By expanding the structural perturbation as $\boldsymbol{\Delta} = \boldsymbol{H}_U\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{H}_V^\top + \boldsymbol{H}_U\boldsymbol{H}_V^\top$, the curvature of the Huber loss in the factor space decomposes into a principal first-order differential component and a higher-order interaction residual. Crucially, within the local basin of attraction guaranteed by our robust spectral initialization, the gradient $\nabla f_{\text{huber}}(\boldsymbol{M})$ is sufficiently small such that the second-order residual is strictly dominated by the positive curvature of the principal component. Substituting the alignment lower bound into our established observation-space curvature directly yields the effective restricted strong convexity for the factor space:

$$\langle \boldsymbol{H}, \nabla^2 f_{\text{huber}}(\boldsymbol{U},\boldsymbol{V})[\boldsymbol{H}]\rangle \ge \frac{q_\tau(1-\delta_{2r})\sigma_r^{*2}}{16}(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2).$$

Hessian of the Regularization Term. In this section, we provide the rigorous derivation of the local curvature generated by the balancing regularizer. The regularization function is defined as:

$$f_{\text{reg}}(\boldsymbol{U}, \boldsymbol{V}) = \frac{\lambda}{8}\|\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{V}^\top \boldsymbol{V}\|_F^2.$$

To facilitate the matrix calculus, we introduce the balancing error matrix $\boldsymbol{G} \triangleq \boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{V}^\top \boldsymbol{V}$. The objective can then be rewritten using the trace inner product as $f_{\text{reg}} = \frac{\lambda}{8}\text{Tr}(\boldsymbol{G}^\top \boldsymbol{G})$.

We compute the differential of $\boldsymbol{G}$ with respect to arbitrary perturbations $d\boldsymbol{U}$ and $d\boldsymbol{V}$:

$$d\boldsymbol{G} = (d\boldsymbol{U})^\top \boldsymbol{U} + \boldsymbol{U}^\top (d\boldsymbol{U}) - (d\boldsymbol{V})^\top \boldsymbol{V} - \boldsymbol{V}^\top (d\boldsymbol{V}).$$

Applying the chain rule, the first variation of the regularization term is:

$$df_{\text{reg}} = \frac{\lambda}{4}\text{Tr}(\boldsymbol{G}^\top d\boldsymbol{G}).$$

The Hessian quadratic form evaluated along the error direction $\boldsymbol{H} = (\boldsymbol{H}_U, \boldsymbol{H}_V)$ corresponds to the second-order differential $d^2 f_{\text{reg}}$, obtained by setting $d\boldsymbol{U} = \boldsymbol{H}_U$ and $d\boldsymbol{V} = \boldsymbol{H}_V$ and differentiating df_{reg} again.

Taking the differential of df_{reg} yields:

$$d^2 f_{\text{reg}} = \frac{\lambda}{4}\text{Tr}\left((d\boldsymbol{G})^\top d\boldsymbol{G} + \boldsymbol{G}^\top d^2\boldsymbol{G}\right).$$

Notice that $d\boldsymbol{U}$ and $d\boldsymbol{V}$ are independent variables (i.e., their second differentials are zero). The second differential of $\boldsymbol{G}$ is exactly $d^2\boldsymbol{G} = 2\boldsymbol{H}_U^\top \boldsymbol{H}_U - 2\boldsymbol{H}_V^\top \boldsymbol{H}_V$. Let us define the directional mapping operator $\mathcal{D}(\boldsymbol{H}) \triangleq \boldsymbol{H}_U^\top \boldsymbol{U} + \boldsymbol{U}^\top \boldsymbol{H}_U - (\boldsymbol{H}_V^\top \boldsymbol{V} + \boldsymbol{V}^\top \boldsymbol{H}_V)$. The exact Hessian quadratic form can be explicitly decomposed into two terms:

$$\langle \boldsymbol{H}, \nabla^2 f_{\text{reg}}(\boldsymbol{U}, \boldsymbol{V})[\boldsymbol{H}]\rangle = \underbrace{\frac{\lambda}{4}\|\mathcal{D}(\boldsymbol{H})\|_F^2}_{T_1} + \underbrace{\frac{\lambda}{2}\text{Tr}\left(\boldsymbol{G}(\boldsymbol{H}_U^\top \boldsymbol{H}_U - \boldsymbol{H}_V^\top \boldsymbol{H}_V)\right)}_{T_2}.$$

We now lower-bound the Hessian quadratic form within the local basin of attraction. Due to the spectral initialization, the current iterates $(\boldsymbol{U}, \boldsymbol{V})$ are close to the optimal balanced factors $(\boldsymbol{U}^*, \boldsymbol{V}^*)$.

First, consider the residual term T_2. In the local neighborhood, the factors are approximately balanced, meaning $\|\boldsymbol{G}\|_F = \|\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{V}^\top \boldsymbol{V}\|_F \leq \mathcal{O}(\|\boldsymbol{H}\|_F)$. Consequently, T_2 is a higher-order term whose magnitude is bounded by the cubic error:

$$T_2 \geq -\mathcal{O}(\|\boldsymbol{H}\|_F^3).$$

Next, we establish the strict positivity of the dominant term T_1. The operator $\mathcal{D}(\boldsymbol{H})$ captures the perturbation along the scaling directions. For any matrix $\boldsymbol{X}$ and symmetric perturbation $\boldsymbol{Y}$, standard eigenvalue inequalities dictate that $\|\boldsymbol{X}^\top \boldsymbol{Y} + \boldsymbol{Y}^\top \boldsymbol{X}\|_F^2 \gtrsim \sigma_{\min}^2(\boldsymbol{X})\|\boldsymbol{Y}\|_F^2$. Because the iterates are locally close to the ground truth, the minimum non-zero singular values of both $\boldsymbol{U}$ and $\boldsymbol{V}$ are bounded below by $\sqrt{\sigma_r^*}$. Summing the contributions from both factors in the worst-case alignment, there exists an absolute constant $C > 0$ such that:

$$T_1 \geq C\lambda\sigma_r^* \left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right).$$

Combining the bounds for T_1 and T_2, we conclude that the regularization term provides a strictly positive curvature along all directions in the factor space:

$$\langle \boldsymbol{H}, \nabla^2 f_{\text{reg}}(\boldsymbol{U}, \boldsymbol{V})[\boldsymbol{H}]\rangle \geq C\lambda\sigma_r^* \left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right) - \mathcal{O}(\|\boldsymbol{H}\|_F^3).$$

This completes the derivation of the regularization curvature.

Total Local Curvature. By combining the curvature bounds from the Huber data-fitting term and the balancing regularizer, we arrive at the total Hessian lower bound for the objective $f = f_{\text{huber}} + f_{\text{reg}}$. Within the local basin of attraction, the higher-order terms $\mathcal{O}(\|\boldsymbol{H}\|_F^3)$ are dominated by the quadratic components.We obtain the final regularity condition:

$$\langle \boldsymbol{H}, \nabla^2 f(\boldsymbol{U}, \boldsymbol{V})[\boldsymbol{H}] \rangle \geq \left(\frac{q_\tau(1-\delta_{2r})\sigma_r^{*2}}{16} + C\lambda\sigma_r^* \right) \left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2 \right).$$

This result establishes that the overall optimization landscape possesses a robust, strictly positive curvature C_{curv} in all directions of the factor space. This "bowl-shaped" geometry is the fundamental catalyst for the linear convergence of the subsequent gradient descent iterations.

D.2 Local Smoothness

We will prove that the gradient of the objective function is locally smooth, which means the gradient does not change too rapidly.This is another key property that ensures the gradient descent algorithm can converge with an appropriate step size.For this analysis, we will decompose the total gradient ∇f into two components: the gradient of the Huber loss function ∇f_{huber} and the gradient of the regularization term ∇f_{reg}.

$$\nabla f = \nabla f_{\text{huber}} + \nabla f_{\text{reg}}$$

Using the norm inequality, we can separately bound the gradient norms of these two parts:

$$\|\nabla f\|_F^2 \leq 2\|\nabla f_{\text{huber}}\|_F^2 + 2\|\nabla f_{\text{reg}}\|_F^2$$

Now, let us analyze the gradient of the Huber loss ∇f_{huber}.The expressions for its gradients with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$ are as follows:

$$\nabla_{\boldsymbol{U}} f_{huber} = \frac{1}{p} \boldsymbol{X}_L^T \mathcal{P}_\Omega \left(\rho_\tau' \left(\boldsymbol{X}_L \boldsymbol{U} \boldsymbol{V}^T \boldsymbol{X}_R^T - \boldsymbol{X}_L \boldsymbol{M}^* \boldsymbol{X}_R^T - \boldsymbol{E} \right) \right) \boldsymbol{X}_R \boldsymbol{V}$$

$$\nabla_{\boldsymbol{V}} f_{huber} = \frac{1}{p} \boldsymbol{X}_R^T \mathcal{P}_\Omega \left(\rho_\tau' \left(\boldsymbol{X}_L \boldsymbol{U} \boldsymbol{V}^T \boldsymbol{X}_R^T - \boldsymbol{X}_L \boldsymbol{M}^* \boldsymbol{X}_R^T - \boldsymbol{E} \right) \right)^T \boldsymbol{X}_L \boldsymbol{U}$$

Considering $\|\boldsymbol{V}\|_2^2 \leq \sigma_1^*$, $\|\boldsymbol{U}\|_2^2 \leq \sigma_1^*$, and using norm inequalities and the property that the operator norms $\|\boldsymbol{X}_L^T\|_2^2 = 1$,$\|\boldsymbol{X}_R^T\|_2^2 = 1$, we can bound the norm of the gradient:

$$\begin{aligned}
\|\nabla_{\boldsymbol{U}} f_{huber}\|_F^2 &= \left\| \frac{1}{p} \boldsymbol{X}_L^T \mathcal{P}_\Omega \left(\rho_\tau' \left(\boldsymbol{X}_L \boldsymbol{U} \boldsymbol{V}^T \boldsymbol{X}_R^T - \boldsymbol{X}_L \boldsymbol{M}^* \boldsymbol{X}_R^T - \boldsymbol{E} \right) \right) \boldsymbol{X}_R \boldsymbol{V} \right\|_F^2 \\
&\leq \frac{2}{p^2} \left\| \boldsymbol{X}_L^T \mathcal{P}_\Omega (\boldsymbol{X}_L \Delta \boldsymbol{X}_R^T) \boldsymbol{X}_R \boldsymbol{V} \right\|_F^2 + \frac{2}{p^2} \left\| \boldsymbol{X}_L^T \mathcal{P}_\Omega (E) \boldsymbol{X}_R \boldsymbol{V} \right\|_F^2 \\
&\leq \frac{2}{p^2} \|\boldsymbol{X}_L^T\|_2^2 \cdot \|\mathcal{P}_\Omega (\boldsymbol{X}_L \Delta \boldsymbol{X}_R^T)\|_F^2 \cdot \|\boldsymbol{X}_R\|_2^2 \cdot \|\boldsymbol{V}\|_2^2 + \frac{2}{p^2} \left\| \boldsymbol{X}_L^T \mathcal{P}_\Omega (E) \boldsymbol{X}_R \right\|_2^2 \cdot \|\boldsymbol{V}\|_F^2 \\
&\leq \frac{2\sigma_1^*}{p^2} \left\| \mathcal{P}_\Omega \left(\boldsymbol{X}_L \Delta \boldsymbol{X}_R^T \right) \right\|_F^2 + \frac{2}{p^2} \left\| \boldsymbol{X}_L^T \mathcal{P}_\Omega (E) \boldsymbol{X}_R \right\|_2^2 \cdot \|\boldsymbol{V}\|_F^2
\end{aligned}$$

According to RIP, we know that

$$\frac{1}{p} \left\| \mathcal{P}_\Omega \left(\boldsymbol{X}_L \boldsymbol{\Delta} \boldsymbol{X}_R^\top \right) \right\|_F^2 = \frac{1}{p} \|\mathcal{A}(\boldsymbol{\Delta})\|_F^2 \leq (1+\delta_{2r}) \|\boldsymbol{\Delta}\|_F^2$$

Combining this with the lower bound for

$$\frac{1}{p} \left\| \mathcal{P}_\Omega \left(\boldsymbol{X}_L \boldsymbol{\Delta} \boldsymbol{X}_R^\top \right) \right\|_F^2 \leq 6(1+\delta_{2r}) \sigma_1^{*2} \left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2 \right)$$

Since we finally get

$$\frac{2\sigma_1^*}{p^2} \left\| \mathcal{P}_\Omega \left(\boldsymbol{X}_L \boldsymbol{\Delta} \boldsymbol{X}_R^\top \right) \right\|_F^2 \leq \frac{12(1+\delta_{2r}) \sigma_1^{*3}}{p} \left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2 \right)$$

Bounding the Projected Noise Term. A critical component of the gradient error stems from the observation noise projected onto the feature space. Let us define the projected noise operator as $\boldsymbol{Z} \triangleq \frac{1}{p}\boldsymbol{X}_L^\top \mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{X}_R$. Following the sub-multiplicative property of matrix norms combining the spectral norm and the Frobenius norm, the noise-induced gradient residual we need to bound is exactly:

$$\frac{2}{p^2}\left\|\boldsymbol{X}_L^\top \mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{X}_R\right\|_2^2 \cdot \|\boldsymbol{V}\|_F^2 = 2\|\boldsymbol{Z}\|_2^2 \cdot \|\boldsymbol{V}\|_F^2.$$

We bound these two terms separately. First, for the factor matrix $\boldsymbol{V} \in \mathbb{R}^{n_2 \times r}$, it is intrinsically a low-rank matrix with rank r. Within the local basin of attraction guaranteed by our spectral initialization, its Frobenius norm squared can be strictly bounded by r times its spectral norm squared:

$$\|\boldsymbol{V}\|_F^2 \le r\|\boldsymbol{V}\|_2^2 \le r\sigma_1^*.$$

Second, the term $\|\boldsymbol{Z}\|_2$ is exactly the spectral norm of the pure noise operator. We have already established its high-probability upper bound in our initialization analysis (see Theorem 3.4) via the Matrix Bernstein inequality. Given that the feature matrices possess orthonormal columns, the noise fluctuation is bounded by:

$$\|\boldsymbol{Z}\|_2^2 \le \frac{C\sigma^2(n_1+n_2)\log d}{p},$$

with probability at least $1 - d^{-1}$.

Substituting these two bounds back into the gradient residual, we obtain a tight, rank-dependent upper bound that successfully avoids the ambient dimension dependency:

$$\frac{2}{p^2}\left\|\boldsymbol{X}_L^\top \mathcal{P}_\Omega(\boldsymbol{E})\boldsymbol{X}_R\right\|_2^2 \cdot \|\boldsymbol{V}\|_F^2 \le \frac{C\sigma_1^*\sigma^2 r(n_1+n_2)\log d}{p}$$

Since,

$$\|\nabla_{\boldsymbol{U}} f_{huber}\|_F^2 \le \frac{12(1+\delta_{2r})\sigma_1^{*3}}{p}(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2) + \frac{C\sigma_1^*\sigma^2 r(n_1+n_2)\log d}{p}$$

The gradient with respect to $\boldsymbol{V}$ follows analogously. Therefore, the total norm of the Huber loss gradient satisfies:

$$\begin{aligned}\|\nabla f_{\text{huber}}\|_F^2 &= \|\nabla_{\mathbf{U}} f_{\text{huber}}\|_F^2 + \|\nabla_{\mathbf{V}} f_{\text{huber}}\|_F^2 \\ &\le \frac{24(1+\delta_{2r})\sigma_1^{*3}}{p}\left(\|\mathbf{H}_U\|_F^2 + \|\mathbf{H}_V\|_F^2\right) + \frac{C\sigma_1^*\sigma^2 r(n_1+n_2)\log d}{p}\end{aligned}$$

Next, we analyze the gradient of the regularization term f_{reg}. The expression for its gradient is as follows:

$$\nabla_{\boldsymbol{U}} f_{\text{reg}} = \frac{\lambda}{2}\boldsymbol{U}\left(\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{V}^\top\boldsymbol{V}\right)$$

$$\nabla_{\boldsymbol{V}} f_{\text{reg}} = -\frac{\lambda}{2}\boldsymbol{V}\left(\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{V}^\top\boldsymbol{V}\right)$$

$$\|\nabla f_{\text{reg}}\|_F^2 = \|\nabla_{\boldsymbol{U}} f_{\text{reg}}\|_F^2 + \|\nabla_{\boldsymbol{V}} f_{\text{reg}}\|_F^2$$

We now turn to the analysis of the gradient of the regularization term with respect to $\boldsymbol{U}$.To facilitate the subsequent analysis, let us first establish a bound for $\|\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{V}^\top\boldsymbol{V}\|_F^2$.

$$\begin{aligned}\|\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{V}^\top\boldsymbol{V}\|_F^2 &= \|\boldsymbol{H}_U^\top\boldsymbol{U}^* + \boldsymbol{U}^{*\top}\boldsymbol{H}_U - \boldsymbol{H}_V^\top\boldsymbol{V}^* - \boldsymbol{V}^{*\top}\boldsymbol{H}_V + \boldsymbol{H}_U^\top\boldsymbol{H}_U - \boldsymbol{H}_V^\top\boldsymbol{H}_V\|_F^2 \\ &\le 2\|\boldsymbol{H}_U^\top\boldsymbol{U}^* + \boldsymbol{U}^{*\top}\boldsymbol{H}_U - \boldsymbol{H}_V^\top\boldsymbol{V}^* - \boldsymbol{V}^{*\top}\boldsymbol{H}_V\|_F^2 + 2\|\boldsymbol{H}_U^\top\boldsymbol{H}_U - \boldsymbol{H}_V^\top\boldsymbol{H}_V\|_F^2 \\ &\le 8\sigma_1^{*2}\left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right) + 2\left(\|\boldsymbol{H}_U\|_F^4 + \|\boldsymbol{H}_V\|_F^4\right)\end{aligned}$$

Equipped with the results above, we are now ready to analyze the gradient of the regularization term with respect to $\boldsymbol{U}$.

$$\begin{aligned}\|\nabla_{\boldsymbol{U}} f_{\text{reg}}\|_F^2 &= \frac{\lambda^2}{4}\|(\boldsymbol{U}^* + \boldsymbol{H}_U)(\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{V}^\top\boldsymbol{V})\|_F^2 \\ &\leq \frac{\lambda^2}{4}\|\boldsymbol{U}^* + \boldsymbol{H}_U\|_2^2\|\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{V}^\top\boldsymbol{V}\|_F^2 \\ &\leq 2\sigma_1^{*3}\lambda^2\left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right) + \frac{\lambda^2\sigma_1^*}{2}(\|\boldsymbol{H}_U\|_F^4 + \|\boldsymbol{H}_V\|_F^4)\end{aligned}$$

Similarly, the bound for $\|\nabla_{\boldsymbol{V}} f_{\text{reg}}\|_F^2$ can be obtained.Finally, by combining the bounds on the gradients of the Huber loss and the regularization term, we prove the local smoothness of the total gradient.

$$\begin{aligned}\|\nabla f_{\text{reg}}\|_F^2 &= \|\nabla_{\boldsymbol{U}} f_{\text{reg}}\|_F^2 + \|\nabla_{\boldsymbol{V}} f_{\text{reg}}\|_F^2 \\ &\leq 4\sigma_1^{*3}\lambda^2\left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right) + \lambda^2\sigma_1^*\left(\|\boldsymbol{H}_U\|_F^4 + \|\boldsymbol{H}_V\|_F^4\right)\end{aligned}$$

Combining the gradients of the loss and regularization terms, we finally get:

$$\begin{aligned}\|\nabla f\|_F^2 &\leq 2\|\nabla f_{\text{huber}}\|_F^2 + 2\|\nabla f_{\text{reg}}\|_F^2 \\ &\leq \left(\frac{48(1+\delta_{2r})\sigma_1^{*3}}{p} + 8\sigma_1^{*3}\lambda^2\right)\left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right) \\ &\quad + \frac{C\sigma_1^*\sigma^2 r(n_1+n_2)\log d}{p}\end{aligned}$$

D.3 Convergence Proof

With the two key properties of local strong convexity and local smoothness established, we can now formally prove the linear convergence of the algorithm. The iterative update rule for gradient descent is as follows:

$$\begin{aligned}\boldsymbol{U}_{k+1} &= \boldsymbol{U}_k - \eta\nabla_{\boldsymbol{U}} f(\boldsymbol{U}_k, \boldsymbol{V}_k) \\ \boldsymbol{V}_{k+1} &= \boldsymbol{V}_k - \eta\nabla_{\boldsymbol{V}} f(\boldsymbol{U}_k, \boldsymbol{V}_k)\end{aligned}$$

We now analyze the relationship between the error at step $k+1$, D_{k+1}^2, and that of step k, D_k^2. Substituting the update rule and expanding the expression yields:

$$\begin{aligned}D_{k+1}^2 &= \left\|\boldsymbol{H}_U^{k+1}\right\|_F^2 + \left\|\boldsymbol{H}_V^{k+1}\right\|_F^2 \\ &= \left\|\boldsymbol{H}_U^k - \eta\nabla_{\boldsymbol{U}} f\right\|_F^2 + \left\|\boldsymbol{H}_V^k - \eta\nabla_{\boldsymbol{V}} f\right\|_F^2 \\ &= \left\|\boldsymbol{H}_U^k\right\|_F^2 - 2\eta\langle\nabla_{\boldsymbol{U}} f, \boldsymbol{H}_U^k\rangle + \eta^2\left\|\nabla_{\boldsymbol{U}} f\right\|_F^2 + \left\|\boldsymbol{H}_V^k\right\|_F^2 - 2\eta\langle\nabla_{\boldsymbol{V}} f, \boldsymbol{H}_V^k\rangle + \eta^2\left\|\nabla_{\boldsymbol{V}} f\right\|_F^2 \\ &= \left\|\boldsymbol{H}_U^k\right\|_F^2 + \left\|\boldsymbol{H}_V^k\right\|_F^2 - 2\eta\left(\left\langle\nabla_{\boldsymbol{U}} f, \boldsymbol{H}_U^k\right\rangle + \left\langle\nabla_{\boldsymbol{V}} f, \boldsymbol{H}_V^k\right\rangle\right) + \eta^2\left(\left\|\nabla_{\boldsymbol{U}} f\right\|_F^2 + \left\|\nabla_{\boldsymbol{V}} f\right\|_F^2\right) \\ &= D_k^2 - 2\eta\left(\left\langle\nabla_{\boldsymbol{U}} f, \boldsymbol{H}_U^k\right\rangle + \left\langle\nabla_{\boldsymbol{V}} f, \boldsymbol{H}_V^k\right\rangle\right) + \eta^2\left(\left\|\nabla_{\boldsymbol{U}} f\right\|_F^2 + \left\|\nabla_{\boldsymbol{V}} f\right\|_F^2\right)\end{aligned}$$

According to local strong convexity,

$$\langle\nabla_{\boldsymbol{U}} f, \boldsymbol{H}_U\rangle + \langle\nabla_{\boldsymbol{V}} f, \boldsymbol{H}_V\rangle \geq \left(\frac{q_\tau(1-\delta_{2r})\sigma_r^{*2}}{16} + C\lambda\sigma_r^*\right)\left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right)$$

According to local smoothness,

$$\begin{aligned}\|\nabla_{\boldsymbol{U}} f\|_F^2 + \|\nabla_{\boldsymbol{V}} f\|_F^2 \leq &\left(\frac{48(1+\delta_{2r})\sigma_1^{*3}}{p} + 8\sigma_1^{*3}\lambda^2\right)\left(\|\boldsymbol{H}_U\|_F^2 + \|\boldsymbol{H}_V\|_F^2\right) \\ &+ \frac{C\sigma_1^*\sigma^2 r(n_1+n_2)\log d}{p}\end{aligned}$$

Next, we apply the two core properties we established previously.We leverage local strong convexity to bound the inner product term and local smoothness to bound the gradient norm term, which yields the following:

$$\begin{aligned} D_{k+1}^2 \leq D_k^2 - 2\eta \left(\frac{q_\tau(1-\delta_{2r})\sigma_r^{*2}}{16} + C\lambda\sigma_r^* \right) D_k^2 \\ + \eta^2 \left(\frac{48(1+\delta_{2r})\sigma_1^{*3}}{p} + 8\sigma_1^{*3}\lambda^2 \right) D_k^2 \\ + \frac{C\sigma_1^*\sigma^2 r(n_1+n_2)\log d}{p} \end{aligned}$$

Provided that we select an appropriate learning rate η to ensure the convergence factor $\rho = (1 - 2\eta c_1 + \eta^2 c_2) < 1$ ($\eta < \frac{p(1-\delta_{2r})\sigma_r^{*2}}{48(1+\delta_{2r})\sigma_1^{*3}+8\sigma_1^{*3}\lambda^2}$), the error is guaranteed to shrink geometrically with each iteration, which demonstrates the algorithm's linear convergence.

E Proof of Corollary 3.8

In this section, we provide the formal derivation of the Euclidean estimation error $\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{M}^*\|_F$ from the recursive alignment error D_k^2 established in Theorem 3.6.

From the recursive bound $D_{k+1}^2 \leq \rho D_k^2 + \mathcal{C}_{\text{stat}}$, as $k \to \infty$, the alignment error D_k^2 converges to a steady-state value D_∞^2:

$$D_\infty^2 \leq \frac{\mathcal{C}_{\text{stat}}}{1-\rho}.$$

Recalling from Appendix D.3 that the contraction factor satisfies $1 - \rho \approx \frac{\eta(1-\delta_{2r})\sigma_r^{*2}}{2}$, and substituting the expression for $\mathcal{C}_{\text{stat}} = \frac{C\sigma_1^*\sigma^2 r(n_1+n_2)\log d}{p}$, we obtain:

$$D_\infty^2 \leq \frac{C\sigma_1^*\sigma^2 r(n_1+n_2)\eta\log d}{p\sigma_r^{*2}},$$

where σ_1^* and σ_r^* denote the largest and r-th singular values of $\boldsymbol{M}^*$, respectively.

The matrix estimation error is bounded by the factor alignment error through local linearization:

$$\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{M}^*\|_F^2 \leq 2\sigma_1^* D_\infty^2.$$

Substituting the steady-state factor error D_∞^2 into the above inequality yields:

$$\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{M}^*\|_F^2 \leq \frac{C(\sigma_1^*)^2\sigma^2 r(n_1+n_2)\eta\log d}{p\sigma_r^{*2}}.$$

By selecting a sufficiently small step size $\eta \approx 1/\sigma_1^*$ to ensure stable convergence, and omitting constant factors and the condition number $\kappa = \sigma_1^*/\sigma_r^*$, the error bound simplifies to:

$$\|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{M}^*\|_F \leq C\sigma\sqrt{\frac{r(n_1+n_2)\log d}{p}}.$$

F Additional numerical experiment

CPU Timing. As demonstrated in Table F.1, our proposed RGDIMC exhibits significant computational advantages across all noise scenarios. Notably, in the presence of heavy-tailed mixture noise, RGDIMC achieves a computational speedup of nearly two orders of magnitude compared to GNIMC and AltMin. While conventional methods suffer from a heavy computational burden when adapting to non-Gaussian outliers, our algorithm maintains a consistently lean execution time, empirically validating our linear complexity analysis. This efficiency, coupled with the superior recovery accuracy discussed previously, underscores the practical scalability of our method for high-dimensional robust matrix completion.

Table F.1: Average CPU Running Time (s) under Different Noise Settings.

Noise Type	**GNIMC**	**Altmin**	**RGDIMC (Ours)**
Gaussian	388.85	920.03	**14.93**
t-distribution	1471.12	823.47	**16.09**
Mixture	1453.15	875.32	**17.26**

PSNG and SSIM. We have incorporated both the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR) into the revised manuscript to provide a more rigorous and perception-aligned evaluation of our image inpainting performance. While PSNR captures absolute pixel-level reconstruction fidelity, SSIM evaluates the preservation of geometric structures and texture homogeneity, aligning more closely with human visual perception. As demonstrated in Table F.2, SSIM effectively reflects the structural preservation superiority of our approach. For instance, consider the highly textured "House" image under extreme mixture noise, where the raw observation yields an initial SSIM of 0.4898. Standard least-squares baselines completely collapse under these extreme outliers, degrading to a poor score of 0.1629. In stark contrast, our robust method successfully neutralizes extreme gradients and systematically restores the underlying structure, achieving the highest SSIM of 0.5733.While the low-rank constraint naturally induces slight high-frequency smoothing, our robust algorithm fundamentally prevents the catastrophic structural collapse suffered by standard non-robust methods.

Table F.2: Comprehensive Quantitative Evaluation across different image characteristics and heavy-tailed noise settings.

		Image: Peppers			**Image: House**		
Noise Type	**Metric**	Noisy Obs.	AltMin	**RGDIMC**	Noisy Obs.	AltMin	**RGDIMC**
Mixture	PSNR	16.25	12.33	**20.82**	16.41	10.81	**18.05**
	SSIM	0.6693	0.4360	**0.8477**	0.4898	0.1629	**0.5733**
Student's t	PSNR	22.33	20.95	**21.73**	22.14	18.46	**20.42**
	SSIM	0.8764	0.8513	**0.8789**	0.7607	0.5879	**0.6967**
Gaussian	PSNR	26.17	22.58	**23.40**	26.04	20.83	**22.15**
	SSIM	0.9415	0.8954	**0.9413**	0.8793	0.7200	**0.7694**

Computational Complexity and Convergence Rate. The proposed RGDIMC framework is highly efficient and scalable to large-scale inductive systems.

(i) Per-Iteration and Space Complexity: In Phase 1, the robust spectral initialization computes a rank-r SVD in $\mathcal{O}(|\Omega|n_1n_2 + n_1n_2r)$ flops. In Phase 2, by carefully ordering matrix multiplication-scomputing projected factors $\tilde{\boldsymbol{U}} = \boldsymbol{X}_L\boldsymbol{U}$ and $\tilde{\boldsymbol{V}} = \boldsymbol{X}_R\boldsymbol{V}$ first and evaluating residuals strictly on observed entries $\Omega$the per-iteration time complexity is $\mathcal{O}\big(|\Omega|r + (d_1n_1 + d_2n_2)r\big)$. The memory footprint is bounded by $\mathcal{O}(|\Omega| + d_1n_1 + d_2n_2)$, entirely avoiding dense $d_1 \times d_2$ operations.

(ii) Total Time Complexity: Based on the local linear convergence in Theorem 3.6 with contraction factor $\rho = 1 - \mathcal{O}(\eta\sigma_r^*) \in (0, 1)$, RGDIMC reaches the statistical error floor $\mathcal{C}_{\text{stat}}$ in $T = \mathcal{O}\left(\frac{1}{\eta\sigma_r^*}\log\left(\frac{D_0^2}{\mathcal{C}_{\text{stat}}}\right)\right)$ iterations, where D_0^2 denotes the initial error from Phase 1. Consequently, the total computational complexity to achieve minimax-optimal robust recovery is bounded by:

$$\text{Total Complexity} = \mathcal{O}\left(|\Omega|n_1n_2 + n_1n_2r + T\big(|\Omega|r + (d_1n_1 + d_2n_2)r\big)\right).$$

This confirms that RGDIMC achieves optimal statistical accuracy in logarithmic iterations while maintaining linear scalability with respect to ambient dimensions and sample size.

Cold Start. We further evaluate the inductive generalization of RGDIMC in a "cold-start" scenario, a common challenge in large-scale recommender systems. In this setup, 20% of users are entirely excluded from the training phase, requiring the model to predict preferences for these unseen users solely via their demographic metadata $\boldsymbol{X}_L$. As illustrated in Figure F.1, the proposed

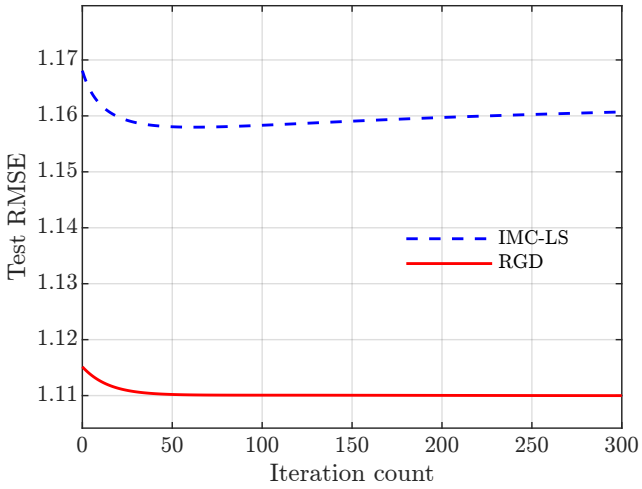


Figure F.1: Robustness under malicious attacks. RGDIMC exhibits rapid convergence and superior stability compared to IMC-LS against 20% malicious users with $\sigma = 30$ noise.

RGDIMC consistently achieves lower prediction errors for unseen users compared to IMC-LS. By effectively suppressing the impact of outlier ratings during training, our robust framework captures a more reliable mapping from the side information space to the latent preference space, thereby enhancing its predictive power for new users.

Initialization sensitivity analysis. We conducted an additional experiment focusing on Phase 1 (Robust Spectral Initialization). We evaluated the relative initialization error $\frac{\|\boldsymbol{M}_0 - \boldsymbol{M}^*\|_F}{\|\boldsymbol{M}^*\|_F}$ where $\boldsymbol{M}_0 = \boldsymbol{U}_0 \boldsymbol{V}_0^\top$ denotes the initial matrix obtained from Phase 1, for different truncation thresholds across three distinct noise distributions: Gaussian, Student-t (df=3), and GMM (Section 4). For each threshold, the evaluation is performed under a fixed random seed to ensure full reproducibility.

As shown in Figure F.2, the initialization error exhibits a clear U-shaped behavior with respect to τ. An overly aggressive threshold removes informative signal components, while an excessively large τ diminishes the truncation effect, reducing the method to vanilla spectral initialization which is vulnerable to heavy-tailed outliers. Notably, the error remains small and stable across a wide basin around the theoretical scaling specified in Theorem 3.4. This empirical robust regime confirms our theoretical finding that τ only needs to be a constant-order quantity rather than an exactly tuned value, which inherently justifies the effectiveness of standard cross-validation in practice.

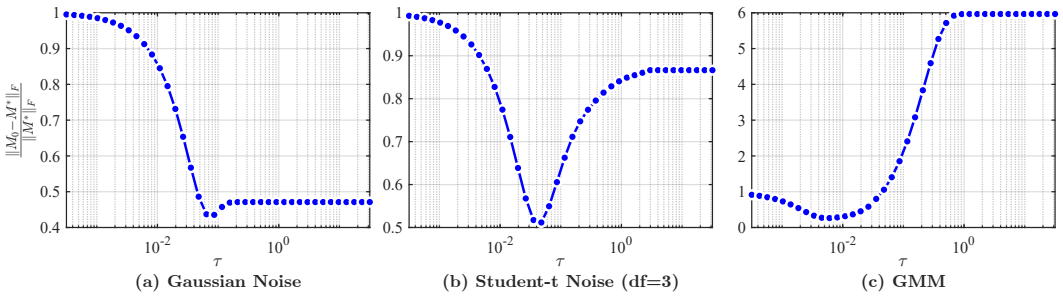


Figure F.2: Sensitivity of robust initialization to truncation thresholds across three distinct noise distributions.

Robustness to Inexact Side Information. To further evaluate the practical applicability of our proposed RGDIMC algorithm, we investigate its performance when the provided side information is inexact, a common challenge in real-world matrix completion tasks.

Following the subspace perturbation framework discussed in Yang and Ma [2026], we introduce a controlled inaccuracy level $\delta \in [0, 0.2]$ to the side information matrices. Specifically, we construct the inexact features by applying a uniform subspace rotation to the ground truth feature matrices, ensuring that the sine of the principal angles between the true and observed subspaces strictly equals δ. Notably, unlike standard perturbation models that apply varying degrees of rotation across different dimensions Yang and Ma [2026], our experimental setup applies this maximum perturbation δ uniformly across all principal directions of the feature subspace. This represents a stricter, worst-case scenario. The observations are simultaneously corrupted by heavy-tailed noise to test the algorithm's dual robustness.

Figure F.3 presents the relative recovery error, defined as $\|\widehat{\mathbf{L}} - \mathbf{L}^*\|_F / \|\mathbf{L}^*\|_F$, as a function of the Side Information Inaccuracy Level (δ). The empirical results demonstrate a highly robust behavior. In the ideal case with exact side information ($\delta = 0$), the RGDIMC algorithm successfully recovers the underlying matrix with a minimal relative error of approximately 0.007. As the inaccuracy level increases, the relative error exhibits a graceful degradation rather than a catastrophic failure. Even under a severe subspace perturbation of $\delta = 0.2$, the relative error only increases to approximately 0.044.

These findings confirm that the proposed algorithm efficiently restricts the negative impact of feature misalignment. By maintaining a highly controlled error bound under both severe heavy-tailed noise and significantly inexact side information, the RGDIMC method proves to be highly reliable for robust inductive matrix completion.

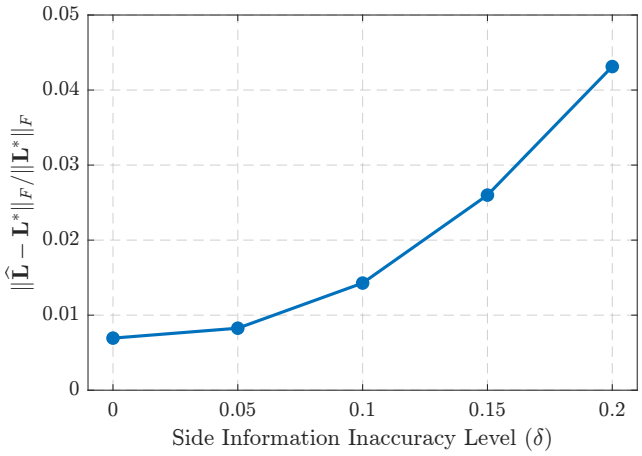


Figure F.3: Relative recovery error $\|\widehat{\mathbf{L}} - \mathbf{L}^*\|_F / \|\mathbf{L}^*\|_F$ of the proposed RGDIMC algorithm against varying levels of side information inaccuracy (δ).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All main claims strictly match the theoretical derivations and empirical results presented in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our algorithm, including the assumptions required for the statistical inference, in the concluding section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Full statements of assumptions and complete proofs for the algorithm's convergence are provided in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide complete MATLAB source code, random seed settings, and detailed descriptions of data generation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The MATLAB source code and scripts to generate all experimental data are included in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental setting and details are thoroughly detailed in the experiments section and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our empirical results report the average metrics over 50 independent trials to ensure statistical significance.

   Guidelines:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Hardware specifications and execution times for the MATLAB experiments are documented in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We briefly discuss the potential societal impacts of robust matrix completion algorithms in the Appendix.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The proposed theoretical algorithm does not involve releasing high-risk models or sensitive datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All third-party assets and baseline algorithms are properly cited and their respective licenses are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release any new datasets; all experiments utilize synthetic data generation or publicly available benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research involves purely theoretical and algorithmic derivations and does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve human subjects, thus IRB approval is not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.