Improving the Optimization Landscape of Matrix Completion with ϵ-close Surrogates

Ziye Ma
Department of Computer Science
City University of Hong Kong
ziyema@cityu.edu.hk

Abstract

Non-convex optimization is central to machine learning and AI, yet unlike its convex counterpart, it still lacks a general and principled framework. In this work, we try to highlight the importance of combining noisy perturbations with over-parametrization mathematically under the classic setting of matrix completion (MC). Matrix completion is an important non-convex recovery problem in which only a subset of a matrix's entries is observed, and the remaining entries must be recovered using low-rank constraints. This problem is notoriously difficult because observations are limited. However, we show that by injecting ϵ-level perturbations into the nullspace of the observation mask, we can construct noisy matrix sensing surrogates whose solutions remain $\mathcal{O}(\epsilon)$-close to the ground truth with valid restricted isometry property (RIP) constants. The existence of RIP further enables powerful tensor frameworks to be applied with theoretical guarantees. This approach reflects a broader empirical trend in which stochasticity or noise improves the curvature of the optimization landscape, allowing over-parameterized (a.k.a large) models to better separate signal from noise. Assuming each matrix entry is observed independently, we establish quantitative recovery guarantees without explicit requirements on sampling rate or incoherence, and discuss how such assumptions can further strengthen our results.

1 Introduction

Unlike convex problems, non-convex optimization landscapes are characterized by multiple local minima, saddle points, and regions of flat curvature, complicating the search for global optima (Jain and Kar, 2017). Consequently, a more in-depth and comprehensive study of such landscapes is essential. One popular way of doing so is to focus on mathematically structured non-convex problems like low-rank matrix recovery. Such problems have garnered increased attention due to their potential to provide deep insights. Low-rank matrix tasks, including matrix completion (MC) and matrix sensing (MS), are important problems since their applications span from collaborative filtering in recommendation systems (Koren et al., 2009), motion detection (Fattahi and Sojoudi, 2020), power system state estimation (Zhang et al., 2017; Jin et al., 2019), to image recovery (Gu et al., 2014) and biomedical imaging (Lustig et al., 2008). More significantly, as these frameworks can encapsulate any polynomial optimization problem (Molybog et al., 2020) and are equivalent to training two-layer quadratic neural networks (Li et al., 2018), their theoretical impact extends well beyond their direct applications in the broader machine learning community.

Matrix sensing generally involves recovering a matrix from a set of linear measurements, formulated as:

$$\min_{X \in \mathbb{R}^{n \times r_{\text{search}}}} \frac{1}{2}\|\mathcal{A}(XX^T) - b\|_2^2 \coloneqq f(XX^T) \quad \text{(MS)} \tag{1}$$

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

For the sake of convenience, we also denote $h(X) \coloneqq f(XX^\top)$. Here, $\mathcal{A}$ acts on the low-rank matrix XX^T (rank bounded by r_{search} by construct) and compares it to a vector of observations $b = \mathcal{A}(M^*)$, with M^* being the rank-r ground truth matrix of interest. $\mathcal{A}(\cdot) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ is a linear function defined as $\mathcal{A}(M) = [\langle A_1, M\rangle, \ldots, \langle A_m, M\rangle]^T$ where $\langle A_i, M\rangle \coloneqq \operatorname{tr}(A_i^\top M)$ and the sensing matrices $\{A_i\}_{i=1}^m$ are given. For simplicity, we assume $r_{\text{search}} = r$. The matrix completion challenge, a special case of matrix sensing, is given by:

$$\min_{X\in\mathbb{R}^{n\times r}} \frac{1}{2}\|\mathcal{A}_\Omega(XX^T - M^*)\|_2^2 \quad \text{(MC)} \tag{2}$$

where $\Omega \subseteq [n] \times [n]$ represents the observed entries of an $n \times n$ matrix. We use the notation N_Ω to denote the matrix where $(N_\Omega)_{i,j} = N_{i,j} \cdot \mathbf{1}_{(i,j)\in\Omega}$ for any arbitrary $N \in \mathbb{R}^{n\times n}$. $\mathcal{A}_\Omega(\cdot)$ is used to specifically denote the sensing operator of the matrix completion problem, where $\mathcal{A}_\Omega(M) = \operatorname{vec}(M_\Omega)$. Matrix completion is distinguished by its reliance on the sample rate and the matrix's incoherence parameters. These parameters dictate the spread of matrix information across its entries and singular vectors (Candès and Tao, 2010). This dependency complicates matrix completion compared to matrix sensing, where challenges are often more tractable due to properties like Restricted Isometry Property (RIP) (see definition B.1). Below, we introduce RSS and RSC (equivalent to RIP) instead for more notational flexibility:

Definition 1.1 (Restricted Strong Smoothness (RSS) and Restricted Strong Convexity (RSC))**.** Function f defined in (1) satisfies the (L_s, r)-RSS property and the (α_s, r)-RSC property if

$$f(M) - f(N) \le \langle M - N, \nabla f(N)\rangle + \frac{L_s}{2}\|M - N\|_F^2$$
$$f(M) - f(N) \ge \langle M - N, \nabla f(N)\rangle + \frac{\alpha_s}{2}\|M - N\|_F^2$$

are satisfied, respectively for all $M, N \in \mathbb{R}^{n\times n}$ with $\operatorname{rank}(M), \operatorname{rank}(N) \le r$. Note that RSS and RSC provide a more expressible way to represent the RIP property Recht et al. (2010), with $\delta_r = (L_s - \alpha_s)/(L_s + \alpha_s)$ after rescaling the sensing operator. For this quadratic loss, $f(M) - f(N) - \langle M - N, \nabla f(N)\rangle = \frac{1}{2}\|\mathcal{A}(M - N)\|_2^2$. Thus the two conditions bound the measurement energy of admissible matrix differences from above and below.

Matrix completion problems, on the other hand, all have missing entries, which means that they could not have a valid RSC constant since the null-space of $\mathcal{A}_\Omega(\cdot)$ is always non-trivial. This is why incoherence condition was proposed to use as a metric to guarantee recovery of completion problems:

Definition 1.2 (μ_0-incoherence)**.** (Ge et al., 2017) Given a rank-r matrix $M \in \mathbb{R}^{n_1\times n_2}$, we say it is μ_0-incoherent if its truncated SVD decomposition $U\Sigma V^\top$ satisfies

$$\|e_i^\top U\|_2 \le \sqrt{\mu_0 r/n_1}, \;\; \|e_j^\top V\|_2 \le \sqrt{\mu_0 r/n_2}$$

$\forall i, j \in [n_1], [n_2]$, where e_i is the i-th standard basis of $\mathbb{R}^{n_1}$ and e_j is the j-th standard basis of $\mathbb{R}^{n_2}$.

Since μ_0 is hard to gauge prior to solving this problem, applying matrix completion guarantees can be more challenging than matrix sensing problems with valid RIP (therefore RSC and RSS) constants.

1.1 Related Works

Recovery Guarantees

The foundational work by Candès and Recht (2009) established that exact matrix recovery is possible from few entries, requiring a sample size of $\mu_0 n^{1.2} r \log(n)$ for $n \times n$ matrices of rank r with incoherence parameter μ_0. Enhancements in recovery guarantees and computational efficiencies followed, including spectral-gradient descent algorithms by Keshavan et al. (2010) and deeper insights into incoherence by Candès and Tao (2010). Studies by Recht (2011) and Gross (2011) expanded on these by demonstrating successful recovery without uniform random sampling, while Ding and Chen (2020) refined sampling orders further to $\mu_0 r \log(\mu_0 r) n \log(n)$. Another notable approach for the classic matrix completion problem (2) is the non-convex Burer-Monteiro (BM) factorized formulation (Ge et al., 2016, 2017; Du et al., 2017), generally requiring $p \ge \text{poly}(\kappa, r, \mu_0, \log n)/n$, with recent works like Zilber and Nadler (2022a) offering progressively sharper bounds under this framework. Other formulations include the convex matrix lasso that uses nuclear norm penalization

Approach	Conditions for recovery w.h.p.
Semi-definite Relaxation (Ding and Chen, 2020)	$p \geq \mathcal{O}\left(\mu_0 r \log(\mu_0 r) n \log(n)/n^2\right)$
BM Factorization (Zilber and Nadler, 2022a)	$p \geq \mathcal{O}\left(\mu_0 r \max\{\log n,\, \mu_0 r \kappa^2\}/n\right)$
Nuclear Norm Lasso (Chen et al., 2020)	$p \geq \mathcal{O}\left(\mu_0^2 \log(n)^3/n\right)$
Spectral Methods (Chen et al., 2021)	$p \geq \mathcal{O}\left(\kappa^2 \mu_0 r \log(n)/n\right)$
Ours	Localized identifiability and shell regularity

Table 1: Brief summary of state-of-the-art recovery guarantees of different approaches in solving (2). This assumes each entry of M^* is sampled uniformly with probability p. μ_0 and κ denotes incoherence and condition number for M^* respectively.

(Chen et al., 2020) and elementary spectral methods (Chen et al., 2021). If we assume that each entry of the desired ground truth matrix M^* is sampled independently with probability p, we could then summarize the different requirements for successful recovery with high probability for these techniques in Table 1. Note that we only included state-of-the-art recovery guarantees. From a high-level, existing requirements of p are only described up to an unspecified constant, so the true p needed for successful recovery is hard to gauge. Our bounds instead express recovery through localized identifiability and shell regularity, complementing these sampling-based guarantees.

Additionally, we introduce some other popular techniques for matrix completion, including greedy algorithms (Lee and Bresler, 2010; Wang et al., 2014), alternating minimization (Haldar and Hernando, 2009; Tanner and Wei, 2016; Wen et al., 2012), iteratively reweighted least squares(IRLS) methods (Mohan and Fazel, 2012; Radhakrishnan et al., 2025), iterative thresholding Klopp (2015) and Riemannian optimization (Mishra et al., 2014; Dai and Milenkovic, 2010)—reviewed comprehensively in Nguyen et al. (2019). Another interesting line of work converts inductive matrix completion into regular MC (Ghassemi et al., 2018; Zilber and Nadler, 2022b), but they have extra information matrices A, B not considered in (2), so it's hard to compare this work to the classic setting.

Why Over-parametrization is Important

Recent studies have highlighted over-parametrization as a crucial strategy in matrix sensing when RIP constants δ_{2r} are suboptimal (i.e., $\delta_{2r} \geq 1/2$). Research by Zhang (2021, 2022) examined cases where the search rank r_{search} exceeds the true rank r, thus increasing the problem's parametrization. Zhang (2022) demonstrated that for $r_{\text{search}} > r[(1+\delta_n)/(1-\delta_n) - 1]^2/4$ and $r \leq r_{\text{search}} < n$, each solution $\hat{X}$ satisfies $\hat{X}\hat{X}^\top = M^*$. Similarly, Ma and Fattahi (2022) established analogous results under RIP-type conditions for the ℓ_1 loss. In addition to the classic over-parametrization of increasing r_{search}, convex relaxation (Recht et al., 2010) is certainly one popular approach by increasing parameters from $\mathcal{O}(nr)$ to $\mathcal{O}(n^2)$. Yalcin et al. (2023) showed that the RIP threshold for exact recovery using SDP can approach 1 when M^* has a high true rank, thus underscoring the efficacy of over-parametrization. Nevertheless, the practical applicability of these conditions is limited, leading Ma et al. (2023) to explore an alternative approach to over-parametrization by lifting the search space of (1) into general tensor space and bank on important concepts from Sums-of-Squares optimization (Parrilo, 2003; Lasserre, 2001) to convert spurious local minimizers of (1) into strict saddle points in the lifted space so that they could be escaped by modern optimizers. Despite its utility in resolving spurious solutions, this tensor approach's applicability to matrix completion remains constrained by the need for a valid RIP constant.

1.2 Notation

Scalar values such as $\sigma_i(M)$ and $\lambda_i(M)$ represent the i-th largest singular value and eigenvalue of matrix M, respectively. The inner product between two matrices $\langle A, B\rangle$ is defined as $\mathrm{tr}(A^\top B)$. The Euclidean norm of a vector v is denoted as $\|v\|$, while $\|M\|_F$ and $\|M\|_2$ are used for the Frobenius and induced l_2 norms of a matrix M; $\|M\|_\infty := \max_{i,j} |M_{ij}|$ denotes its entrywise maximum norm. Vectorization $\mathrm{vec}(M)$ stacks the columns of M into a vector, where $\mathrm{mat}(v)$ is its reverse operation. The set of integers from 1 to n is expressed as $[n]$. Moreover, $\circ l$ indicates a repeated Cartesian product for l times, $\oslash$ refers to the Kronecker product, and $\otimes$ signifies the tensor outer product. If these

notations come with subscripts, they denote the dimension along which the operation is performed. Finally, if $S \subseteq [n] \times [m]$ represents a subset of indices of a $n \times m$ matrix, then N_S refers to the masked matrix with entries N_{ij} on S and zeros elsewhere, and S^{c} denotes its complement.

Throughout this paper, $M \succeq 0$ denotes a symmetric positive semidefinite matrix. The factorization $M = XX^\top$, with $X \in \mathbb{R}^{n\times r}$, restricts the candidate matrix to this cone with rank at most r. Orthogonal transformations XQ represent the same matrix, so we measure recovery error between $XX^\top$ and M^*. For matrix completion, $\mathcal{A}_\Omega(M) = \mathrm{vec}(M_\Omega)$ retains all n^2 coordinates, placing zeros outside Ω, so $\|\mathcal{A}_\Omega(M)\|_2 = \|M_\Omega\|_F$. The complementary masks satisfy $M = M_\Omega + M_{\Omega^c}$ and $\|M\|_F^2 = \|M_\Omega\|_F^2 + \|M_{\Omega^c}\|_F^2$. Thus the original and perturbed operators use observation vectors of the same dimension. We use $\widetilde{M} = \widetilde{X}\widetilde{X}^\top$ for a global minimizer of the surrogate, distinguishing it from the ground truth M^*.

For convenience, $x^{\otimes l}$ denotes the repeated tensor outer product of a vector x, with entries $[x^{\otimes l}]_{i_1\dots i_l} = x_{i_1}\cdots x_{i_l}$. For matrices, the Kronecker product arranges its output as a block matrix, while the tensor outer product retains the separate modes. In particular, for nonzero X, $\mathrm{vec}(X)^{\otimes l} \in \mathbb{R}^{nr\cdots l}$ is a symmetric rank-one tensor even when X has r linearly independent columns. Hence rank-one in the lifted formulation refers to tensor rank, while the recovered matrix $XX^\top$ may have rank r. Tensor contractions sum over the modes indicated by their subscripts; Appendix B provides their coordinate definitions.

2 The Perturbed Matrix Completion Formulation

As explained above, most literature regarding the recovery guarantees of matrix sensing problems require some valid RIP (RSC and RSS) constant. A positive RSC constant on the whole matrix space implies a trivial nullspace, which is impossible for incomplete entry observations. To demonstrate why this is, let's consider a 2×2 matrix recovery problem, and say we observed three entries of some $M \in \mathbb{R}^{2\times 2}$ except for the lower-right entry. This will correspond to the case where

$$\mathcal{A}_\Omega(M) = \mathrm{vec}\left(\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix} \odot M\right) = \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0\end{bmatrix} \mathrm{vec}(M) \coloneqq T_\Omega\,\mathrm{vec}(M) \in \mathbb{R}^4 \tag{3}$$

For this $\mathcal{A}_\Omega$ to exhibit any RIP constant $\delta < 1$, it is required that $\|\mathcal{A}_\Omega(M)\|_2^2 \geq (1-\delta)\|M\|_F^2$, meaning that the T_Ω matrix above cannot output 0 unless M is a zero matrix. Nevertheless, for this specific example, even if we observed three out of four entries, we can simply set $\mathrm{vec}(M) = [0,0,0,1]^\top$ to make $\mathcal{A}_\Omega(M) = 0$, violating the RIP condition. This is a simple example showing us that RIP condition will not hold for matrix completion problems unless all entries are observed, regardless of its size. Therefore, it begs the question of *whether we could use the better studied, more powerful over-parametrized MS framework to offer guarantees for MC problems?* In this work, we give a positive answer to this question via the following three steps:

1. **Surrogate Construction:** We construct a surrogate problem (ϵ-MC) to solve by slightly changing $\mathcal{A}_\Omega$. ϵ-MC problem is a noisy MS problem that has more favorable theoretical properties.
2. **Surrogate Accuracy:** We prove that the surrogate's global solution approximates the true matrix M^* with high probability under mild conditions.
3. **Tensor Framework Adaptation:** To address high RIP constants in the perturbed MS setting, we extend and adapt the lifted tensor framework from Ma et al. (2024), leveraging over-parametrization to achieve recovery.

Despite the lack of RIP constant, a surprisingly simple solution exists. The primary issue is the zero entries in the diagonal of T_Ω, contributing to a non-trivial nullspace. By perturbing these zero entries slightly with a small number $\epsilon \in (0,1]$, we can eliminate the nullspace. Revisiting (3), consider a perturbed sensing operator $\mathcal{A}_{\Omega,\epsilon} = \mathrm{diag}([1,1,1,\epsilon])\,\mathrm{vec}(M) \coloneqq T_{\Omega,\epsilon}\,\mathrm{vec}(M)$, in which $T_{\Omega,\epsilon}$ has a trivial nullspace as promised. However, different operators lead to different observations. For example, when considering the case of (3), using $\mathcal{A}_\Omega$ and $\mathcal{A}_{\Omega,\epsilon}$ on the same matrix results in:

$$\mathcal{A}_\Omega(M) = [M_{1,1}M_{1,2}M_{2,1}0]^\top \longrightarrow \mathcal{A}_{\Omega,\epsilon}(M) = [M_{1,1}M_{1,2}M_{2,1}\epsilon M_{2,2}]^\top$$

Therefore, the original observation b can be considered a noisy observation under the context of $\mathcal{A}_{\Omega,\epsilon}$, with

$$b = \mathcal{A}_{\Omega,\epsilon}(M^*) + w_\epsilon,\;\; w_\epsilon = \begin{bmatrix}0 & 0 & 0 & -\epsilon M^*_{2,2}\end{bmatrix}^\top \tag{4}$$

where $w_\epsilon \in \mathbb{R}^{n^2}$ can be considered a noise term. With this idea in place, we formally introduce our perturbed MC problem to solve:

$$\min_{X \in \mathbb{R}^{n \times r_{\text{search}}}} \frac{1}{2}\|\mathcal{A}_{\Omega,\epsilon}(XX^T) - b\|_2^2 = \frac{1}{2}\|\mathcal{A}_\Omega(XX^T - M^\star)\|_{\text{F}}^2 + \frac{\varepsilon^2}{2}\|\mathcal{A}_{\Omega^c}(XX^T)\|_{\text{F}}^2 \quad (\epsilon\text{-MC}) \tag{5}$$

The optimizer receives only Ω and the observed entries $b = \mathcal{A}_\Omega(M^\star)$, with zeros at missing coordinates. The complement term in (5) regularizes the candidate matrix and uses no missing ground-truth values. The identity $w_\epsilon = -\epsilon\mathcal{A}_{\Omega^c}(M^\star)$ is used only in analysis: its unknown entries cancel those of $\mathcal{A}_{\Omega,\epsilon}(M^\star)$, leaving the same observed data b.

Equivalently, (5) is a weighted low-rank approximation (Razenshteyn et al., 2016; Li et al., 2016) with a deterministic weight floor on missing entries. Our contribution is the perturbation and its surrogate-accuracy certificate for symmetric PSD factors and a two-level mask; extending it to general weights and asymmetric factors is left for future work.

We denote the ϵ-MC objective as $f_{w_\epsilon}(XX^\top)$, and Ω^{c} denotes the complement set to Ω. Essentially, we're transforming a noiseless matrix completion problem into a noisy matrix sensing problem with operator $\mathcal{A}_{\Omega,\epsilon}$ and deterministic noise w_ϵ. This approach, notably, ensures the attainment of valid RSS/RSC parameters, equivalent to RIP constants.

Lemma 2.1 (RIP/RSC of ϵ-MC)**.** *For $\varepsilon \in (0, 1]$, the ε-MC objective satisfies $(1, n)$-RSS and (ε^2, n)-RSC:*

$$\varepsilon^2\|M\|_{\text{F}}^2 \le \|\mathcal{A}_{\Omega,\varepsilon}(M)\|_2^2 \le \|M\|_{\text{F}}^2 \quad \textit{for all } M \in \mathbb{R}^{n\times n}.$$

If entries are observed independently with probability p, then for every fixed $M \neq 0$,

$$\mathbb{E}\|\mathcal{A}_{\Omega,\varepsilon}(M)\|_2^2 = [p + \varepsilon^2(1-p)]\|M\|_{\text{F}}^2.$$

For $\delta \in (0, 1)$, put $u = \log(1/\delta)$ and

$$t_\delta(M) = \sqrt{\frac{2p(1-p)u}{s_\infty(M)}} + \frac{2u}{3s_\infty(M)}, \qquad s_\infty(M) = \frac{\|M\|_{\text{F}}^2}{\|M\|_\infty^2}.$$

With probability at least $1 - \delta$,

$$\|\mathcal{A}_{\Omega,\varepsilon}(M)\|_2^2 \ge \underbrace{\left[\varepsilon^2 + (1-\varepsilon^2)\max\{0, p - t_\delta(M)\}\right]}_{\alpha_{p,\varepsilon}^2(\delta,M)}\|M\|_{\text{F}}^2. \tag{6}$$

Equation (6) *describes the energy of a fixed matrix; Section 3 develops the uniform shell estimates used for recovery.*

The global RSC constant ε^2 is tight in general. Independent observation can improve the energy lower bound for a fixed sufficiently spread matrix, as (6) shows. For the perturbation $\epsilon = 5 \times 10^{-5}$ used in our experiments, the deterministic constants are $\alpha_s = 2.5 \times 10^{-9}$ and $L_s = 1$, giving condition ratio 4×10^8 and normalized RIP constant approximately 0.999999995. Thus the perturbation provides positive curvature in the matrix variable, while the factorized objective can retain spurious solutions. This motivates applying a lifted framework that can handle large RIP constants. The proof of Lemma 2.1 appears in Appendix C.

3 Quality of the Surrogate ϵ-MC Problem

Now we have established that ϵ-MC problems have valid RSC constants, but another major question to ask is that *whether the surrogate problem solves the original one*? Generally speaking this statement is false because $\|\mathcal{A}_{\Omega,\epsilon}(M^*) - b\|_2^2 \neq 0$; however, if we impose some simple technical constraints, then it's possible to show that the new global optimum of (5) stays $\mathcal{O}(\epsilon)$ close to M^*. Please find all proofs and supporting lemmas in Appendix C, and we start with the notion of identifiability.

3.1 Identifiability of Matrix Completion Problems

For the unperturbed matrix-completion problem, define the feasible completion set

$$\mathcal{S}_0(\Omega, M^\star; r) := \{M \succeq 0 : \ \text{rank}(M) \le r, \ \mathcal{A}_\Omega(M) = \mathcal{A}_\Omega(M^\star)\}. \tag{7}$$

Unique completability means $\mathcal{S}_0(\Omega, M^\star; r) = \{M^\star\}$. To quantify separation from $M^\star$, set $D := \sqrt{5}\|M^\star\|_{\mathrm{F}}$ and define, for $0 < \rho \le D$,

$$\mu_{\mathrm{id},D}(\rho) := \inf_{\substack{M \succeq 0,\ \mathrm{rank}(M) \le r \\ \rho \le \|M - M^\star\|_{\mathrm{F}} \le D}} \frac{\|\mathcal{A}_\Omega(M - M^\star)\|_{\mathrm{F}}^2}{\|M - M^\star\|_{\mathrm{F}}^2}. \tag{8}$$

The PSD, rank, and distance constraints define a compact annulus, on which the ratio is continuous. The upper radius follows from global optimality, rather than an imposed search constraint: writing $b_\perp := \|\mathcal{A}_{\Omega^c} M^\star\|_{\mathrm{F}}$, comparison of any global minimizer $\widetilde{M}$ with $M^\star$ gives

$$\|\mathcal{A}_\Omega(\widetilde{M} - M^\star)\|_{\mathrm{F}}^2 + \varepsilon^2 \|\mathcal{A}_{\Omega^c}\widetilde{M}\|_{\mathrm{F}}^2 \le \varepsilon^2 b_\perp^2.$$

Thus mask orthogonality and the triangle inequality yield

$$\|\widetilde{M} - M^\star\|_{\mathrm{F}}^2 \le (\varepsilon^2 + 4) b_\perp^2 \le 5\|M^\star\|_{\mathrm{F}}^2 = D^2,$$

using $0 < \varepsilon \le 1$. Here $b_\perp$ is a scalar, distinct from the observation vector b.

Lemma 3.1 (Bounded identifiability)**.** *$M^\star$ is the unique member of $\mathcal{S}_0(\Omega, M^\star; r)$ within distance D if and only if $\mu_{\mathrm{id},D}(\rho) > 0$ for every $0 < \rho \le D$. In particular, global unique completability is sufficient.*

Theorem 2 of Pimentel-Alarcón et al. (2016) gives explicit combinatorial conditions on the observation pattern under which completion is unique for almost every rank-r matrix. The implication for ϵ-MC is as follows.

Lemma 3.2 (Small ϵ guarantees recovery under identifiability)**.** *Fix $0 < \rho \le D$. If*

$$\mu_{\mathrm{id},D}(\rho)\rho^2 > \varepsilon^2 \|\mathcal{A}_{\Omega^c} M^\star\|_{\mathrm{F}}^2, \tag{9}$$

then every global minimizer of (5) *satisfies $\|\widetilde{M} - M^\star\|_{\mathrm{F}} < \rho$.*

If $b_\perp = 0$, the level-set bound already gives $\widetilde{M} = M^\star$; if $\rho > D$, localization is automatic. Lemma C.2 establishes monotonicity and a lower bound for the bounded modulus under its stated sharpness assumption. To control the missing energy and obtain explicit error bounds, we next impose more structure on Ω.

3.2 Beyond Identifiability Under Bernoulli Sampling of Entries

Arguably, the most common instances of MC problems is one where each entry is observed independently with probability p. In other words, each entry of the mask $\mathcal{A}_\Omega$ is a Bernoulli random variable. Under this classic setting, we can significantly strengthen the bound between M^* and the global solution $\widetilde{M}$ of (5) using a probabilistic argument. Specifically, for $\rho > 0$, let

$$\mathcal{H}(\rho) := \{H = M - M^\star :\ M \succeq 0,\ \mathrm{rank}(M) \le r,\ \|H\|_{\mathrm{F}} \le \rho\}. \tag{10}$$

Lemma 3.2 says if $\mu_{\mathrm{id},D}(\rho)\rho^2$ dominates the perturbation bias, then global optimality of (5) forces the actual error $\Delta = \widetilde{M} - M^\star$ into $\mathcal{H}(\rho)$. A uniform Bernstein argument on this fixed shell then gives the main Bernoulli guarantee.

Theorem 3.3 (Main result for quality of $\widetilde{M}$)**.** *Assume $\mathcal{A}_\Omega \sim \mathrm{Ber}(p)$ entrywise, let $\widetilde{X}$ be a global minimizer of* (5)*, and set $\widetilde{M} = \widetilde{X}\widetilde{X}^T$. Fix $0 < \rho \le D$ and $\chi \in (0, \min\{p, 1-p\})$. Assume ϵ is chosen with* [1]*:*

$$\mu_{\mathrm{id},D}(\rho)\rho^2 > \varepsilon^2(1 - p + \chi)\|M^\star\|_{\mathrm{F}}^2. \tag{11}$$

then

$$\|\widetilde{M} - M^\star\|_{\mathrm{F}}^2 \le \frac{1 - p + \chi}{p - \chi}\,\varepsilon^2 \|M^\star\|_{\mathrm{F}}^2 \tag{12}$$

holds with probability at least $1 - \delta_{\textit{bernoulli}}$, where $\delta_{\textit{bernoulli}} =$

$$2\exp\left[-c\min\left\{\frac{\chi^2 s_4(M^\star)}{p(1-p)},\ \chi s_\infty(M^\star)\right\}\right] + 2\exp\left[(4n+2)r\log\left(\frac{36}{\chi}\right) - c\min\left\{\frac{\chi^2 s_{4,\rho}}{p(1-p)},\ \chi s_{\infty,\rho}\right\}\right]. \tag{13}$$

for some universal constant c, where $s_4(A) := \|A\|_{\mathrm{F}}^4/\|A\|_4^4$, $s_\infty(A) := \|A\|_{\mathrm{F}}^2/\|A\|_\infty^2$ for any non-zero $A \in \mathbb{R}^{n\times n}$ and $s_{4,\rho} :- \inf_{H\in\mathcal{H}(\rho)\setminus\{0\}} s_4(H)$, $s_{\infty,\rho} :- \inf_{H\in\mathcal{H}(\rho)\setminus\{0\}} s_\infty(H)$.

[1] The bounded-modulus gap is an additional hypothesis; the radius lies in $(0, D]$.

Although the theorem has no explicit hard cut-off for p, decreasing p worsens both the error bound and its success probability, which can become vacuous. The bounded-modulus gap and the shell parameters $s_{4,\rho}, s_{\infty,\rho}$ are substantive assumptions; Bernoulli sampling alone does not guarantee them. If H has s nonzero entries of equal magnitude, then $s_4(H) = s_\infty(H) = s$, illustrating their dependence on how the error is spread across entries. Incoherence of $M^\star$ and a sufficiently small radius allow explicit lower bounds: Proposition C.6 gives $s_{4,\rho}, s_{\infty,\rho} \geq n/(16\mu_0 r)$. The probability bound remains conservative.

3.3 How traditional "goodness" in MC improves our results

Theorem 3.3 gives an $\mathcal{O}(\varepsilon)$ surrogate bound under its localization and shell regularity conditions. However, the probability (13) with which it holds is quite conservative to respect its generality. Nevertheless, if standard assumptions in the literature are invoked, especially an order-of-magnitude lower-bound on p and that the ground truth matrix M^* satisfies the incoherence condition as per Definition 1.2, then our results can be further sharpened under the same framework.

The most recent recovery result (Zilber and Nadler, 2022a) under the BM factorized objective (2) shows that if $M^\star$ is rank r, μ_0-incoherent, and has condition number κ, then under

$$np \geq \mathcal{O}\left(\frac{1}{\eta^2}\,\mu_0 r \max\left\{\log n,\ \frac{\mu_0 r\kappa^2}{\eta^2}\right\}\right), \tag{14}$$

for some $\eta < 1$, the sampling operator is uniformly well conditioned on a local incoherent neighborhood of $M^\star$. We show in Lemma C.8 that if we further constrain $\rho < \rho_{\mathrm{GNMR}}$, where

$$\rho_{\mathrm{GNMR}} := \min\left\{\frac{\eta\,\sigma_r(M^\star)}{c_e}, \frac{\sigma_r(M^\star)}{2}, \frac{c_\mu\sqrt{\mu_0 r/n}}{1 + c_\mu\sqrt{\mu_0 r/n}}\,\sigma_r(M^\star)\right\}, \qquad c_\mu := \frac{\sqrt{3/2}-1}{\sqrt{2}}. \tag{15}$$

and assume that (14) holds, then lower-bound to $\mathcal{A}_\Omega(\widetilde{M} - M^*)$ can be improved to $\mathcal{O}(1-\eta)$ with much higher probability. Combine this with Theorem 3.3 yields the main result of this subsection:

Theorem 3.4 (Main theorem under GNMR local RIP)**.** *Fix $\eta \in (0,1)$ and $\chi \in (0, 1-p)$. Assume: (i) $M^\star$ is rank r, μ_0-incoherent, and has condition number κ; (ii) the sample size condition* (14) *holds; (iii) $0 < \rho \leq \rho_{\mathrm{GNMR}} \leq \sigma_r(M^\star)/2 < D$; (iv) $\mu_{\mathrm{id},D}(\rho)\rho^2 > \varepsilon^2(1-p+\chi)\|M^\star\|_{\mathrm{F}}^2$. Then every global minimizer of* (5) *satisfies*

$$\|\widetilde{M} - M^\star\|_{\mathrm{F}}^2 \leq \frac{1-p+\chi}{p(1-\eta)^2}\,\varepsilon^2\|M^\star\|_{\mathrm{F}}^2 \tag{16}$$

with probability at least $1 - \delta_{GNMR}$ with

$$\delta_{GNMR} = \frac{3}{n^3} + 2\exp\left[-c\min\left\{\frac{\chi^2 s_4(M^\star)}{p(1-p)},\ \chi s_\infty(M^\star)\right\}\right]. \tag{17}$$

This result complements the local guarantees of Zilber and Nadler (2022a): it bounds the accuracy of a global surrogate minimizer once the bounded-modulus gap places it in the local-RIP neighborhood. Here the perturbation controls the accuracy of a global surrogate solution, while the over-parametrization in the next section addresses how to reach an accurate solution.

4 The Computational Solution via Over-parametrization

Now that we are able to reformulate the original MC problem (2) into the new ϵ-MC problem (5), it presents us with a new challenge. Although now (5) admits valid RSC/RSS constants, this is nevertheless still a difficult matrix sensing problem to solve due to its small RSC constant (or large RIP constant). Thus, it is important that we apply an over-parametrized framework to deal with it in order to compensate for the poor geometric uniformity.

To this end, we employ the lifted tensor framework proposed in Ma et al. (2023), since it has the ability to deal with arbitrarily small RSC constants. However, in their original work, measurements were assumed to be clean, and this is incompatible with our framework. Thus, in this section we generalize the original results in Ma et al. (2023) and Ma et al. (2024) to demonstrate how the inclusion of noise could affect guarantees when using a higher-order tensor parametrization.

First of all, we present the lifted tensor problem when our observations are corrupted by some random noise $\tilde{w} \in \mathbb{R}^m$,

$$\min_{\mathbf{w}\in\mathbb{R}^{nro l}} \quad \|\langle \mathbf{A}^{\otimes l}, \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\mathbf{w})\rangle_{2*[l]}\rangle - \tilde{b}^{\otimes l}\|_F^2 \tag{18}$$

where $\tilde{b} = \mathcal{A}(M^*)+\tilde{w}$, and $\mathbf{A} \in \mathbb{R}^{m\times n\times n}$ is a three-way tensor which can be seen as a concatenation of all sensing matrices $\{A_i\}_{i=1}^m$, and $\mathbf{w} \in \mathbb{R}^{nro l}$ is the tensor decision variable used to increase the parametrization of $X \in \mathbb{R}^{n\times r}$. Here, $^{\otimes l}$ simply denotes l times of repeated tensor outer product, and $\mathbf{P}$ is just another constant permutation tensor used for correct multiplication. The gist of this paper is not on the tensor formulation, thus many details are deferred to Appendix B, and interested readers can learn more about general tensor knowledge and problem details there. For convenience's sake, we define $f^l(\cdot) : \mathbb{R}^{n\circ 2l} \mapsto \mathbb{R}$ and $h^l(\cdot) : \mathbb{R}^{[n\times r]\circ l} \mapsto \mathbb{R}$ as $f^l(\mathbf{M}) \coloneqq \|\langle \mathbf{A}^{\otimes l}, \mathbf{M}\rangle - \tilde{b}^{\otimes l}\|_F^2$ and $h^l(\mathbf{w}) = f^l(\langle \mathbf{w}, \mathbf{w}\rangle_{2*[l]})$, with $\nabla f^l(\cdot) = \nabla_{\mathbf{M}} f^l(\cdot)$ and $\nabla h^l(\cdot) = \nabla_{\mathbf{w}} h^l(\cdot)$.

In the original works, it was proven that the lifted formulation (18) is able to convert spurious second-order points in (1) to strict saddle points (critical points with local escape direction) via its drastic over-parametrization if this spurious solution is somehow far away from the ground truth M^*. Therefore, we have established a new result catering to our noisy formulation (5) below to show that even in the noisy situation, similar results still hold:

Theorem 4.1. *Consider an arbitrary second-order point $\hat{X} \in \mathbb{R}^{n\times r}$ of the factorized matrix sensing objective in the form of* (1) *where its observations b could be potentially corrupted by some random noise $\tilde{w} \in \mathbb{R}^m$ (i.e. $b = \tilde{b}$). Assuming that the linear operator $\mathcal{A}(\cdot)$ in* (1) *satisfies the RSC and RSS conditions with constants α_s, L_s respectively. Then $\hat{\mathbf{w}} = \mathrm{vec}(\hat{X})^{\otimes l}$ is a strict saddle of* (18) *with a rank-1 symmetric escape direction if*

$$\|M^* - \hat{X}\hat{X}^\top\|_F^2 \geq \frac{L_s}{\alpha_s}\lambda_r(\hat{X}\hat{X}^\top)\,\mathrm{tr}(M^*) + \frac{\|\tilde{w}\|_2^2}{\alpha_s} \tag{19}$$

with an odd l satisfying

$$l > \frac{1}{1-\log_2(2\beta)}, \quad \beta \coloneqq \frac{L_s\,\mathrm{tr}(M^*)\lambda_r(\hat{X}\hat{X}^\top)}{\alpha_s\|M^* - \hat{X}\hat{X}^\top\|_F^2 - \|\tilde{w}\|_2^2}. \tag{20}$$

The proof of this theorem is located in Appendix C. The theorem highlights how the conversion radius from spurious solutions to strict saddles is influenced by the norm of the noise $\tilde{w}$. Setting $\tilde{w} = 0$ allows this theorem to coincide with Theorem 4 from Ma et al. (2024). While the above result holds for all noisy matrix sensing problems, if we plug-in the configurations of our ϵ-MC formulation with $\tilde{w} = -\epsilon\mathcal{A}_{\Omega^c}M^\star$, we get an immediate corollary:

Corollary 4.2. *Consider the matrix completion problem of completing a $n \times n$, rank-r matrix M^*, where $\Omega \subseteq [n] \times [n]$ denotes the set of observed entries and Ω^c denotes the unobserved entries. If we solve its ϵ-MC surrogate as per* (5)*, then any rank-1 local minima $\hat{\mathbf{w}} = \mathrm{vec}(\hat{X})^{\otimes l}$ of* (18) *with $\mathcal{A} = \mathcal{A}_{\Omega,\epsilon}$ satisfies:*

$$\frac{\|M^* - \hat{X}\hat{X}^\top\|_F}{\lambda_r(\hat{X})} < \frac{\sqrt{\mathrm{tr}(M^*)}}{\alpha_{p,\epsilon}} + \frac{\|\mathcal{A}_{\Omega^c}M^\star\|_{\mathrm{F}}}{\lambda_r(\hat{X})} \tag{21}$$

Here $\alpha_{p,\epsilon} \geq \epsilon$ denotes an admissible norm-scale lower bound for the comparisons required by Theorem 4.1, with quadratic RSC constant $\alpha_s = \alpha_{p,\epsilon}^2$. The deterministic choice $\alpha_{p,\epsilon} = \epsilon$ is always valid; a sharper uniform bound requires separate justification beyond the fixed-matrix estimate (6)*.*

Although this results gives a direct upper-bound to the normalized distance between the local minimum $\hat{X}$ and M^*, the term $\|\mathcal{A}_{\Omega^c}M^\star\|_{\mathrm{F}}/\lambda_r(\hat{X})$ is unknown and could be potentially very large. To address this issue, we present the following main result by assuming more specific structures:

Theorem 4.3 (Main Result)**.** *Consider the setting described in Corollary 4.2 where Ω is obtained with i.i.d. observation of probability p. We choose an ϵ that satisfies* (11) *for some $0 < \rho \leq D$. Then applying the lifted formulation to the surrogate ϵ-MC problem yields the following results:*

For any rank-1 local minima $\hat{\mathbf{w}} = \mathrm{vec}(\hat{X})^{\otimes l}$ of (18)*, if the conditions of* (20) *are met with $L_s = 1, \alpha_s = \alpha_{p,\epsilon}^2$ and $\tilde{w} = 0$, then it is guaranteed to satisfy that:*

$$\|M^* - \hat{X}\hat{X}^\top\|_F < \frac{\lambda_r(\hat{X})}{\alpha_{p,\epsilon}}\sqrt{\mathrm{tr}(M^*)} + \varepsilon\sqrt{\frac{1-p+\chi}{p-\chi}}\|M^\star\|_{\mathrm{F}} \tag{22}$$

for any $\chi \in (0, \min\{p, 1-p\})$ with probability at least $1 - \delta_{bernoulli}$ given in (13)*. Additionally, if the ground truth matrix M^* is also μ_0-incoherent with some constant μ_0, the bound can be further strengthened to:*

$$\|M^* - \hat{X}\hat{X}^\top\|_F < \frac{\lambda_r(\hat{X})}{\alpha_{p,\epsilon}}\sqrt{\mathrm{tr}(M^*)} + \varepsilon\frac{\sqrt{(1-p+\chi)/p}\|M^\star\|_{\mathrm{F}}}{1-\eta} \tag{23}$$

by imposing the requirements on p and ϵ in Theorem 3.4, where the probability also improves to $\delta \geq \delta_{GNMR}$ in (17)*.*

Theorem 4.3 follows by applying the noiseless lifting comparison of Ma et al. (2024) around the surrogate minimizer $\widetilde{M}$: global optimality makes its objective gap nonpositive, without requiring zero measurement residual. Theorems 3.3 and 3.4 bound $\|\widetilde{M} - M^\star\|_{\mathrm{F}}$, so the triangle inequality gives (22) and (23). The noisy results in Theorem 4.1 and Corollary 4.2 compare directly to $M^\star$ and are not used here.

Although the result above holds only for rank-one critical points, its approximately rank-one extension is discussed in Theorem B.10 of Appendix B. Theorem B.9 establishes approximately rank-one behavior for plain GD under small-initialization and finite-horizon conditions. Appendix E gives the escape-curvature constant and a conditional perturbed-GD consequence: complete strict-saddle alternatives, derivative bounds, and trajectory containment yield $\widetilde{\mathcal{O}}(L_h\Delta_h/\eta_g^2)$ iterations to the protected neighborhood (Jin et al., 2017). These additional conditions are distinct from plain GD's implicit bias; a full GD–Tensor PCA runtime bound remains open.

5 Numerical Experiments

In this section, we evaluate our method[2] against both classic and state-of-the-art solvers on two families of matrix completion (MC) instances. In addition to the uniformly and independently sampled instances as in Theorem 3.3, we also include the hard instance from Yalçın et al. (2022), which is known to admit exponentially many spurious solutions:

$$\Omega = \{(i,i), (i,2k), (2k,i) \mid \forall i \in [n],\ k \in [\lfloor n/2 \rfloor]\}. \tag{24}$$

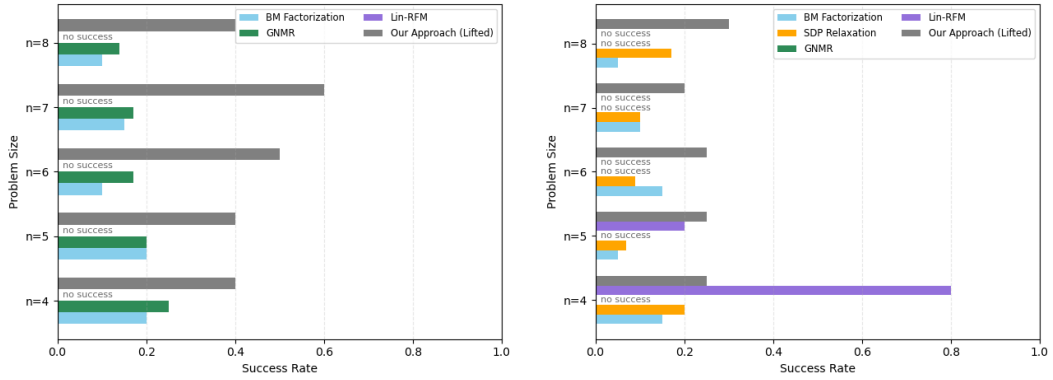
We compare our methods to: (i) the classic Burer–Monteiro (BM) factorization baseline for (2); (ii) the SDP relaxation (Candès and Tao, 2010); (iii) GNMR (Zilber and Nadler, 2022a), a state-of-the-art BM method that solves a linearized least-squares subproblem each iteration (the corrected reference implementation, Algorithm 3 with $\alpha = -1$, is evaluated separately in Table 2); (iv) Lin–RFM (Radhakrishnan et al., 2025), an SVD-free IRLS/Schatten-p spectral-reweighting scheme; and (v) our lifted approach with $l = 3$. Our method produces a tensor $\mathbf{w}_T$ after T gradient steps, and Tensor PCA (Ma et al., 2024) extracts a principal component $X_T \in \mathbb{R}^{n\times r}$ via $\mathbf{w}_T \approx \mathrm{vec}(X_T)^{\otimes l}$, with X_T being returned for (2). We declare success if $\|X_T X_T^\top - M^*\|_F \leq 0.05$ and report *success rates* over 20 trials per problem, rather than average reconstruction error, because they are robust to outliers and better capture instances where ϵ-MC may be an invalid surrogate. Ablation studies of ϵ and p are included in Appendix A.

We see from figure 1 that across both instance families and sizes n, our lifted approach attains consistently higher success rates. The GNMR baseline in the original figure is competitive on the structured instance but degrades on the independent model, while Lin–RFM, as an IRLS/Schatten-p method, shows strong performance at smaller sizes in the independent model and is less effective on the structured instance. Appendix D examines unsuccessful recoveries on the structured instance: all ten runs fit observed entries to mean-squared error below 5×10^{-8}, while seven have missing-entry error about 2.67. The error maps locate these failures on unobserved entries.

To separate perturbation from lifting, we compare direct optimization of MC and ϵ-MC on rank-five $n = 200$ instances. Each solver uses 10 matched problems, masks, and initializations per sampling rate, with $\epsilon = 5 \times 10^{-5}$. Success means Frobenius error at most 0.05; for reference GNMR (Algorithm 3, $\alpha = -1$), we evaluate $UV^\top$. Table 2 reports success rates (%) and their change in percentage points.

Here Δ is the ϵ-MC success rate minus the MC success rate. Only two of the 120 paired outcomes change, both for Adam, showing that perturbation alone does not systematically resolve direct solvers'

[2]Code: `https://github.com/MOFA-LAB/mc_as_noisyms`. Experiments run on an M1 Max MacBook Pro.



(a) Problem Instance (24)

(b) Independent Entries with $p = 0.15$

Figure 1: We use $\epsilon = 5 \times 10^{-5}$, a learning rate of 2×10^{-2}, and an initialization scale of $\xi = 10^{-4}$ for (18). SDP was not deployed for (a) since the solution is not unique in the relaxed regime. For (b), the variability of Ω necessitates running 10 distinct problem instances for each size n.

Table 2: Direct recovery at $n = 200$: perturbation changes only Adam's success rate.

	$p = 0.02$			$p = 0.05$			$p = 0.08$		
Solver	MC	ϵ-MC	Δ (pp)	MC	ϵ-MC	Δ (pp)	MC	ϵ-MC	Δ (pp)
Adam	0	0	0	30	40	$+\mathbf{10}$	70	60	$-\mathbf{10}$
GNMR (reference)	0	0	0	0	0	0	0	0	0
GD (Armijo)	0	0	0	90	90	0	100	100	0
L-BFGS	0	0	0	90	90	0	100	100	0

failures. Figure 1 tests the combined approach, whereas Table 2 isolates perturbation within each solver. Together they illustrate how perturbation provides a regular sensing surrogate while lifting addresses the remaining optimization difficulty. Appendix F gives further controls.

Choosing ϵ. Appendix A studies the structured $n = 6$, $l = 3$ instance with 100 trials per perturbation. At the 0.05 threshold, success is approximately 26% at $\epsilon = 5 \times 10^{-4}$ and 36% at 5×10^{-5}, versus about 9% for BM, but drops to 11% at 10^{-5}. This illustrates the tradeoff between perturbation bias and conditioning. The sampling-rate sweep shows the largest improvement over BM at small p.

Limitations. Choosing ϵ involves a tradeoff: smaller perturbations reduce surrogate bias but weaken the guaranteed curvature, making optimization harder. Explicit order-l tensors require $\mathcal{O}((nr)^l)$ storage, limiting our implementation to small-scale problems. Simulated Oracle Direction (SOD) (Shen et al., 2026) uses lifting to construct escape candidates in the original factor space without storing the tensor, alleviating this cost. Appendix G reports favorable results on larger structured instances; full adoption for noisy ϵ-MC remains future work. Our guarantees concern symmetric PSD factors, a two-level mask, and rank-one or approximately rank-one lifted points. Broader settings and complete runtime and recovery guarantees, including Tensor PCA and factor extraction, remain further directions.

6 Conclusion

In conclusion, we shift from the classical goal of exact recovery of M^* to solving a carefully designed surrogate that remains provably close to M^* and whose landscape is amenable to over-parameterization. Our analysis separates surrogate accuracy from the optimization benefits of lifting, and the experiments illustrate how perturbation and over-parametrization can complement one another in matrix completion.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62506314), the Research Grants Council of Hong Kong (Grant No. 21208525), and City University of Hong Kong (Project No. 9382001).

References

Candes, E. J. and Plan, Y. (2010). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *arXiv preprint arXiv:1001.0339*.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080.

Chen, Y., Chi, Y., Fan, J., Ma, C., et al. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121.

Comon, P., Golub, G., Lim, L.-H., and Mourrain, B. (2008). Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279.

Dai, W. and Milenkovic, O. (2010). Set: An algorithm for consistent matrix completion. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3646–3649. IEEE.

Ding, L. and Chen, Y. (2020). Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 66(11):7274–7301.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. (2017). Gradient descent can take exponential time to escape saddle points. *Advances in neural information processing systems*, 30.

Fattahi, S. and Sojoudi, S. (2020). Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis. *Journal of Machine Learning Research*, 21:1–51.

Ge, R., Jin, C., and Zheng, Y. (2017). No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1233–1242.

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

Ghassemi, M., Sarwate, A., and Goela, N. (2018). Global optimality in inductive matrix completion. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2226–2230. IEEE.

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

Gu, S., Zhang, L., Zuo, W., and Feng, X. (2014). Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869.

Ha, W., Liu, H., and Barber, R. F. (2020). An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4):2927–2955.

Haldar, J. P. and Hernando, D. (2009). Rank-constrained solutions to linear matrix equations using powerfactorization. *IEEE Signal Processing Letters*, 16(7):584–587.

Jain, P. and Kar, P. (2017). Non-convex optimization for machine learning. *arXiv preprint arXiv:1712.07897*.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. (2017). How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR.

Jin, M., Molybog, I., Mohammadi-Ghazi, R., and Lavaei, J. (2019). Towards robust and scalable power system state estimation. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3245–3252. IEEE.

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

Klopp, O. (2015). Matrix completion by singular value thresholding: sharp bounds. *Electronic Journal of Statistics*, 9(2):2348–2369.

Kolda, T. G. (2015). Numerical optimization for symmetric tensor decomposition. *Mathematical Programming*, 151(1):225–248.

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

Lasserre, J. B. (2001). Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817.

Lee, K. and Bresler, Y. (2010). Admira: Atomic decomposition for minimum rank approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416.

Li, Q., Zhu, Z., and Tang, G. (2019). The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96.

Li, Y., Liang, Y., and Risteski, A. (2016). Recovery guarantee of weighted low-rank approximation via alternating minimization. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2358–2367. PMLR.

Li, Y., Ma, T., and Zhang, H. (2018). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR.

Lustig, M., Donoho, D. L., Santos, J. M., and Pauly, J. M. (2008). Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72–82.

Ma, J. and Fattahi, S. (2022). Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *arXiv preprint arXiv:2202.08788*.

Ma, Z., Lavaei, J., and Sojoudi, S. (2024). Algorithmic regularization in tensor optimization: Towards a lifted approach in matrix sensing. *Advances in Neural Information Processing Systems*, 36.

Ma, Z., Molybog, I., Lavaei, J., and Sojoudi, S. (2023). Over-parametrization via lifting for low-rank matrix sensing: Conversion of spurious solutions to strict saddle points. In *International Conference on Machine Learning*. PMLR.

Mishra, B., Meyer, G., Bonnabel, S., and Sepulchre, R. (2014). Fixed-rank matrix factorizations and riemannian low-rank optimization. *Computational Statistics*, 29:591–621.

Mohan, K. and Fazel, M. (2012). Iterative reweighted algorithms for matrix rank minimization. *The Journal of Machine Learning Research*, 13(1):3441–3473.

Molybog, I., Madani, R., and Lavaei, J. (2020). Conic optimization for quadratic regression under sparse noise. *The Journal of Machine Learning Research*, 21(1):7994–8029.

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

Parrilo, P. A. (2003). Semidefinite programming relaxations for semialgebraic problems. *Mathematical programming*, 96(2):293–320.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):510.

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

Radhakrishnan, A., Belkin, M., and Drusvyatskiy, D. (2025). Linear recursive feature machines provably recover low-rank matrices. *Proceedings of the National Academy of Sciences*, 122(13):e2411325122.

Razenshteyn, I., Song, Z., and Woodruff, D. P. (2016). Weighted low rank approximations with provable guarantees. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pages 250–263. ACM.

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

Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501.

Shen, T., Yang, J., He, J., Gao, K., Zheng, Z., and Ma, Z. (2026). Escaping local minima provably in non-convex matrix sensing: A deterministic framework via simulated lifting. *arXiv preprint arXiv:2602.05887*.

Tanner, J. and Wei, K. (2016). Low rank matrix completion by alternating steepest descent methods. *Applied and Computational Harmonic Analysis*, 40(2):417–429.

Wang, Z., Lai, M.-J., Lu, Z., Fan, W., Davulcu, H., and Ye, J. (2014). Rank-one matrix pursuit for matrix completion. In *International Conference on Machine Learning*, pages 91–99. PMLR.

Wen, Z., Yin, W., and Zhang, Y. (2012). Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361.

Yalcin, B., Ma, Z., Lavaei, J., and Sojoudi, S. (2023). Semidefinite programming versus burer-monteiro factorization for matrix sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yalçın, B., Zhang, H., Lavaei, J., and Sojoudi, S. (2022). Factorization approach for low-complexity matrix completion problems: Exponential number of spurious solutions and failure of gradient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 319–341. PMLR.

Zhang, H., Bi, Y., and Lavaei, J. (2021). General low-rank matrix optimization: Geometric analysis and sharper bounds. *Advances in Neural Information Processing Systems*, 34:27369–27380.

Zhang, R. Y. (2021). Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime. *arXiv preprint arXiv:2104.10790*.

Zhang, R. Y. (2022). Improved global guarantees for the nonconvex burer–monteiro factorization via rank overparameterization. *arXiv preprint arXiv:2207.01789*.

Zhang, Y., Madani, R., and Lavaei, J. (2017). Conic relaxations for power system state estimation with line measurements. *IEEE Transactions on Control of Network Systems*, 5(3):1193–1205.

Zilber, P. and Nadler, B. (2022a). Gnmr: A provable one-line algorithm for low rank matrix recovery. *SIAM journal on mathematics of data science*, 4(2):909–934.

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

A Additional Ablation Studies

In this section we complement our main experiments with ablation studies on the user-chosen perturbation level ϵ and the sampling probability p for the structured instance in problem (24). Throughout, we consider the setting $n = 6$, depth $\ell = 3$, and initialization magnitude 10^{-4}.

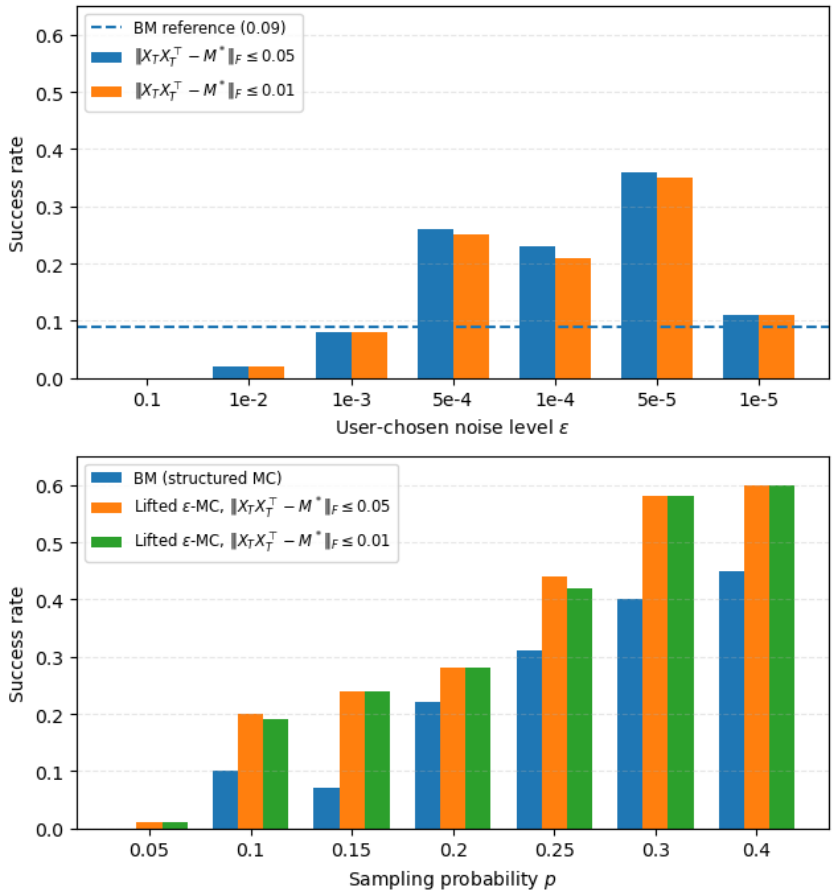


Figure 2: Ablation studies for the lifted ϵ-MC formulation with $n = 6$, $\ell = 3$, initialization magnitude 10^{-4}. *Top:* Success rates with respect to the perturbation level ϵ, under two reconstruction criteria $\|X_T X_T^\top - M^\star\|_F \le 0.05$ and $\|X_T X_T^\top - M^\star\|_F \le 0.01$, together with the BM baseline success rate (dashed line) for the structured instance (24). Success rate is polled from 100 independent trials each. *Bottom:* Success rates of the independently sampled problem with respect to sampling probability p, comparing the BM approach and the lifted ϵ-MC method under both reconstruction criteria. For each p level, 10 random masks are generated and 20 trials each are run for a total of 200 trials.

Ablation on ϵ. For the first study we fix the sampling pattern as in problem (24) and vary the perturbation level $\epsilon \in \{0.1, 10^{-2}, 10^{-3}, 5\times10^{-4}, 10^{-4}, 5\times10^{-5}, 10^{-5}\}$. For each ϵ, we record the empirical success rates under the two reconstruction criteria $\|X_T X_T^\top - M^\star\|_F \le 0.05$ and $\|X_T X_T^\top - M^\star\|_F \le 0.01$. As shown in the top panel of Figure 2, the success curves for the two thresholds closely track each other, indicating that this problem behaves as a "success or failure" scenario rather than exhibiting a gradual degradation in reconstruction quality.

These trends are consistent with the error bound in Theorem 4.3, which decomposes into a *computational part*, proportional to $1/\epsilon$ (this rate is rather conservative and could be improved, please see Lemma 2.1), and an *accuracy part* e_1 capturing the mismatch between the ϵ-perturbed objective and the original matrix completion problem. For large ϵ (e.g., 0.1 and 10^{-2}), the perturbation error e_1 dominates and the success rate is low even if the optimizer reaches a global minimum of the ϵ-MC landscape. As ϵ decreases into an intermediate regime (roughly between 5×10^{-4} and 5×10^{-5}), the

two terms balance and the success rate peaks. When ϵ becomes too small (e.g., 10^{-5}), the $1/\epsilon$ factor renders the problem numerically more challenging, and the empirical success rate deteriorates. In this example, the useful range spans approximately an order of magnitude; ϵ requires tuning, and this sweep does not supply a universal selection rule.

Ablation on sampling probability p. In the second study we fix $\epsilon = 5\times10^{-4}$ and vary the sampling probability $p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4\}$. For each p, we again estimate the success rates over 10 randomly generated masks with 20 trials each amounting to 200 trials in total for both the BM method and our approach. The results are summarized in the bottom panel of Figure 2.

As p increases, both methods become more successful, as expected from having more observed entries. The lifted ϵ-MC approach exhibits a clear advantage in the low-p regime, where the BM formulation struggles to recover the independently sampled instance. At larger p, the BM baseline itself achieves high success rates, which naturally reduces the margin between the two methods. At the same time, further gains for our approach appear limited by the same computational term in the error bound: while the accuracy part improves rapidly with p, the difficulty of optimizing the lifted problem becomes the dominant factor.

Overall, these ablation studies illustrate how the choice of ϵ and the sampling probability p jointly govern the empirical behavior of the lifted ϵ-MC formulation.

B Additional Details for Noisy Lifted Framework

B.1 Additional Definitions

Definition B.1 (RIP)**.** (Candès and Recht, 2009) Given a natural number p, the linear map $\mathcal{A} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ is said to satisfy δ_p-RIP if there is a constant $\delta_p \in [0, 1)$ such that

$$(1 - \delta_p)\|M\|_F^2 \leq \|\mathcal{A}(M)\|^2 \leq (1 + \delta_p)\|M\|_F^2$$

holds for matrices $M \in \mathbb{R}^{n\times n}$ satisfying $\operatorname{rank}(M) \leq p$.

Definition B.2 (Tensor)**.** As a generalization of the way vectors are used to parametrize finite-dimensional vector spaces, we use *arrays* to parametrize tensors generated from product of finite-dimensional vector spaces, as per Comon et al. (2008). In particular, we define an l-way array as such:

$$\mathbf{a} = \{a_{i_1 i_2 \dots i_l} | 1 \leq i_k \leq n_k, 1 \leq k \leq l\} \in \mathbb{R}^{n_1\times\cdots\times n_l}$$

Note that in this paper tensors and arrays can be regarded as synonymous since there exists an isomorphism between them. Moreover, if $n_1 = \cdots = n_l$, then we call this tensor(array) an l-order(way), n-dimensional tensor. For the convenience of tensor representation, we use the notation $\mathbb{R}^{n\circ l}$ with $n \circ l \coloneqq n \times \cdots \times n$. In this work, tensors are denoted with bold variables, and other fonts are reserved for matrices, vectors, and scalars unless specified otherwise.

Definition B.3 (Symmetric Tensor)**.** Similar to the definition of symmetric matrices, for an order-l tensor $\mathbf{a}$ with the same dimensions (i.e., $n_1 = \cdots = n_l$), also called a cubic tensor, it is said that the tensor is symmetric if its entries are invariance under any permutation of their indices:

$$a_{i_{\sigma(1)}\cdots i_{\sigma(l)}} = a_{i_1\cdots i_l} \quad \forall\sigma, \quad i_1, \dots, i_l \in \{1, \dots, n\}$$

where $\sigma \in \mathcal{G}_l$ denotes a specific permutation and $\mathcal{G}_l$ is the symmetric group of permutations on $\{1, \dots, l\}$. We denote the set of symmetric tensors as $\mathrm{S}^l(\mathbb{R}^n)$.

Definition B.4 (Rank of Tensors)**.** The rank of a cubic tensor $\mathbf{a} \in \mathbb{R}^{n\circ l}$ is defined as

$$\operatorname{rank}(\mathbf{a}) = \min\{r | \mathbf{a} = \sum_{i=1}^{r} u_i \otimes v_i \otimes \cdots \otimes w_i\}$$

for some vector $u_i, \dots, w_i \in \mathbb{R}^n$. Furthermore, according to Kolda (2015), if $\mathbf{a}$ is a symmetric tensor, then it can be decomposed as:

$$\mathbf{a} = \sum_{i=1}^{r} \lambda_i u_i \otimes \cdots \otimes u_i \coloneqq \sum_{i=1}^{r} \lambda_i u_i^{\otimes l}$$

and the rank is conveniently defined as the number of nonzero λ_i's, which is very similar to the rank of symmetric matrices indeed. The most important concept in our paper is rank-1 tensors, and for any tensor $\mathbf{a}$, a necessary and sufficient condition for it to be rank-1 is that

$$\mathbf{a} = u^{\otimes l}$$

for some $u \in \mathbb{R}^n$.

Definition B.5 (Tensor Multiplication)**.** Outer product is an operation carried out on a pair of tensors, denoted as $\otimes$. The outer product of 2 tensors $\mathbf{a}$ and $\mathbf{b}$, respectively of orders l and p, is a tensor of order $l + p$, denoted as $\mathbf{c} = \mathbf{a} \otimes \mathbf{b}$ such that:

$$c_{i_1 \dots i_l j_1 \dots j_p} = a_{i_1 \dots i_l} b_{j_1 \dots j_p}$$

When the 2 tensors are of the same dimension, this product is such that $\otimes : \mathbb{R}^{n \circ l} \times \mathbb{R}^{n \circ p} \mapsto \mathbb{R}^{n \circ (l+p)}$. Henceforth, we use the shorthand notation

$$\underbrace{a \otimes \cdots \otimes a}_{l \text{ times}} \coloneqq a^{\otimes l}$$

We also define an inner product of two tensors. The mode-q inner product between the 2 aforementioned tensors having the same q-th dimension is denoted as $\langle \mathbf{a}, \mathbf{b} \rangle_q$. Without loss of generality, assume that $q = 1$ and

$$[\langle \mathbf{a}, \mathbf{b} \rangle_q]_{i_2 \dots i_l j_2 \dots j_p} = \sum_{\alpha=1}^{n_q} a_{\alpha i_2 \dots i_l} b_{\alpha j_2 \dots j_p}$$

Note that when we write $\langle \cdot, \cdot \rangle_q$, we count the q-th dimension of the first entry. Indeed, this definition of inner product can also be trivially extended to multi-mode inner products by just summing over all modes, denoted as $\langle \mathbf{a}, \mathbf{b} \rangle_{q,\dots,s}$.

Lemma B.6 (Section 10.2 Petersen et al. (2008))**.** *For four arbitrary matrices A, B, C, D of compatible dimensions, it holds that*

$$\langle A \otimes B, C \otimes D \rangle_{2,4} = AC \otimes BD \tag{25}$$

Definition B.7 (Variational Eigenvalue of Tensors Ma et al. (2024))**.** For a given tensor $\mathbf{w} \in \mathbb{R}^{n \circ l}$, we define its kth variational eigenvalue (v-Eigenvalue) $\lambda_k^v(\mathbf{w})$ as

$$\lambda_k^v(\mathbf{w}) \coloneqq \max_{\substack{S \\ \dim(S)=k}} \min_{\mathbf{u} \in S} \frac{|\langle \mathbf{w}, \mathbf{u} \rangle|}{\|\mathbf{u}\|_F^2}, \quad k \in [n]$$

where S is a subspace of $\mathbb{R}^{n \circ l}$ that is spanned by a set of orthogonal, symmetric, rank-1 tensors. Its dimension denotes the number of orthogonal tensors that span this space.

B.2 Formulation Details

As noted in our main formulation (18), the decision variable $\mathbf{w}$ is a tensor of dimension $nr \times \cdots \times nr$, since it serves as a repeated outer product of $\text{vec}(X)$ with $X \in \mathbb{R}^{n \times r}$ being our original decision variable in (1) (here we assume $r_{\text{search}} = r$). The permutation $\mathbf{P}$ is needed in order to convert $\mathbf{w} \in \mathbb{R}^{nr \circ l}$ back to $\mathbb{R}^{[n \times r] \circ l}$ in order to do meaningful inner products. $\mathbf{P} \in \mathbb{R}^{n \times r \times nr}$ is defined as

$$\langle \mathbf{P}, \text{vec}(X) \rangle_3 = X \quad \forall X \in \mathbb{R}^{n \times r}, n, r \in \mathbb{Z}^+$$

Such $\mathbf{P}$ can be easily constructed via filling appropriate scalar "1"s in the tensor. Via Lemma B.1, we also know that

$$\langle \mathbf{P}^{\otimes l}, \text{vec}(X)^{\otimes l} \rangle_{3*[l]} = (\langle \mathbf{P}, \text{vec}(X) \rangle_3)^{\otimes l} = X^{\otimes l} \tag{26}$$

Notationally, we abbreviate $\langle \mathbf{P}^{\otimes l}, \mathbf{w} \rangle_{3*[l]}$ as $\mathbf{P}(\mathbf{w})$ for enhanced readability for an arbitrary tensor $\mathbf{w}$ with dimension greater or equal to 2.

Since we also make extensive use of first and second order critical points of (18), we present them here for accessibility:

Lemma B.8. *The tensor $\hat{\mathbf{w}} \in \mathbb{R}^{nr \circ l}$ is an SOP of* (18) *if and only if*

$$\langle \nabla f^l(\langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\hat{\mathbf{w}}) \rangle_{2*[l]}), \mathbf{P}(\hat{\mathbf{w}}) \rangle_{2*[l]} = 0, \tag{27a}$$

$$\begin{aligned} 2\langle \nabla f^l(\langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\hat{\mathbf{w}}) \rangle_{2*[l]}), \langle \mathbf{P}(\Delta), \mathbf{P}(\Delta) \rangle_{2*[l]} + \\ \| \langle \mathbf{A}^{\otimes l}, \langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\Delta) \rangle_{2*[l]} + \langle \mathbf{P}(\Delta), \mathbf{P}(\hat{\mathbf{w}}) \rangle_{2*[l]} \rangle \|_F^2 \geq 0 \quad \forall \Delta \in \mathbb{R}^{nr \circ l} \end{aligned} \tag{27b}$$

with (27b) *being a necessary and sufficient condition for $\hat{\mathbf{w}}$ to be a FOP and $\nabla f_w^l(\mathbf{M})$ is defined as*

$$\nabla f_w^l(\mathbf{M}) = \langle (\mathbf{A}^{\otimes l})^* \mathbf{A}^{\otimes l}, \mathbf{M} \rangle - [\langle \mathbf{A}^* \mathbf{A}, M^* \rangle + \langle \mathbf{A}^*, \tilde{w} \rangle]^{\otimes l} \tag{28}$$

The proof to this lemma is highly technical and can be obtained by slightly changing the proof to Lemma 7 in Ma et al. (2024) by changing $b = \mathcal{A}(M^*)$ to $\tilde{b} = \mathcal{A}(M^*) + \tilde{w}$ defined above.

Here we present the full theorem regarding implicit bias in (18):

Theorem B.9. *Consider a finite-horizon gradient descent trajectory $\{\mathbf{w}_t\}_{t \in [T]}$ of* (18) *with $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla h^l(\mathbf{w}_t)$ starting from the initialization $\mathbf{w}_0 = \xi x_0^{\otimes l}$ with $\xi \in \mathbb{R}$ denoting the scale of the initialization, η representing the step-size and $x_0 \in \mathbb{R}^{nr}$ being an arbitrary vector with $\|x_0\|_2^2 = 1$. Then there exists $t(\kappa, l) \geq 1$ and $\kappa < 1$ such that*

$$\frac{\lambda_2^v(\mathbf{w}_t)}{\lambda_1^v(\mathbf{w}_t)} \leq \kappa, \qquad \forall t \in [t(\kappa, l), T] \tag{29}$$

if the initialization scale ξ is sufficiently small, where $t(\kappa, l)$ is expressed as

$$t(\kappa, l) = \left\lceil \ln \left(\frac{\|x_0\|_2^l}{\kappa |v_1^\top x_0|^l} \right) \ln \left(\frac{1 + \eta \sigma_1^l(U)}{1 + \eta \sigma_2^l(U)} \right)^{-1} \right\rceil \tag{30}$$

where $\sigma_1(U)$ and $\sigma_2(U)$ denote the first and second singular values of U and v_1, v_2 are the associated singular vectors, with

$$U = \langle \mathbf{A}_r, \tilde{b} \rangle_1 \in \mathbb{R}^{nr \times nr}, \quad \mathbf{A}_r \coloneqq I_r \oslash_{2,3} \mathbf{A} \tag{31}$$

Next, we present a technical extension of Theorem 4.1 and Theorem B.9, showing how gradient descent initialized with small scale can help ensure that second-order points of lifted version of (5) remain very close to M^* along the optimization trajectory

Theorem B.10. *Consider a generic matrix completion problem under the same premise as given in Theorem 4.1. Assume that the symmetric tensor $\hat{\mathbf{w}} \in \mathbb{R}^{nr \circ l}$ is a second-order point (local minima) of* (18) *that is κ-rank-1 with $\kappa \leq \mathcal{O}(1/\|M^*\|_F^2)$. This can be achieved by initializing the vanilla gradient algorithm at $\mathbf{w}_0 = \xi x_0^{\otimes l}$ with a sufficiently small $\xi > 0 \in \mathbb{R}$. Then after iterations $t(\kappa, l)$ given in* (30)*, Theorem B.9 ensures that all tensors along the trajectory will become κ-rank-1.*

If $\hat{\mathbf{w}}$'s major spectral decomposition is given as $\hat{\mathbf{w}} = \lambda_S \hat{x}^{\otimes l} + \hat{\mathbf{w}}^\dagger$ with $\hat{x} \in \mathbb{R}^{nr}$ being a FOP of (5) *(ensured by Proposition 2 in Ma et al. (2024)), we know that*

$$\|M^* - \hat{X}\hat{X}^\top\|_F < \frac{1}{\epsilon} \lambda_r(\hat{X}) \sqrt{\operatorname{tr}(M^*)} + \mathcal{O}(\sqrt{r} \kappa^{1/2l}) + e_1 \tag{32}$$

holds with probability at least q, under the condition that l is odd and meets the requirement:

$$l > \frac{1}{1 - \log_2(2\beta)}, \quad \beta \coloneqq \frac{\operatorname{tr}(M^*) \lambda_r(\hat{X}\hat{X}^\top)}{\epsilon^2 \left(\|M^* - \hat{X}\hat{X}^\top\|_F^2 - \mathcal{O}(r\kappa^{1/l}) - e_2 \right)}. \tag{33}$$

where e_1, e_2, and q are identical to those given in Theorem 4.1 depending on different MC instances.

The proof of this Theorem is omitted because it directly follows from Theorem 4.1 , Theorem B.9, and Theorem 2 in Ma et al. (2024).

C Proof Details

C.1 Auxiliary Results

We collect the auxiliary ingredients used in Section 3.

Lemma C.1 (Basic inequality)**.** *Let $\widetilde{M}$ be any global minimizer of* (5)*. Then*

$$\frac{1}{2}\|\mathcal{A}_\Omega \Delta\|_{\mathrm{F}}^2 + \frac{\varepsilon^2}{2}\|\mathcal{A}_{\Omega^{\mathrm{c}}}(\Delta + M^\star)\|_{\mathrm{F}}^2 \le \frac{\varepsilon^2}{2}\|\mathcal{A}_{\Omega^{\mathrm{c}}} M^\star\|_{\mathrm{F}}^2. \tag{34}$$

Consequently,

$$\|\mathcal{A}_\Omega \Delta\|_{\mathrm{F}} \le \varepsilon \|\mathcal{A}_{\Omega^{\mathrm{c}}} M^\star\|_{\mathrm{F}}, \tag{35}$$

$$\|\mathcal{A}_{\Omega^{\mathrm{c}}} \widetilde{M}\|_{\mathrm{F}} \le \|\mathcal{A}_{\Omega^{\mathrm{c}}} M^\star\|_{\mathrm{F}}. \tag{36}$$

Moreover, every global minimizer satisfies

$$\|\Delta\|_{\mathrm{F}}^2 \le (\varepsilon^2 + 4)\|\mathcal{A}_{\Omega^{\mathrm{c}}} M^\star\|_{\mathrm{F}}^2 \le D^2. \tag{37}$$

Lemma C.2 (Monotonicity and lower bound for the bounded identifiability modulus)**.** *The map $\rho \mapsto \mu_{\mathrm{id},D}(\rho)$ is nondecreasing on $(0, D]$. Assume there exist $\alpha_\star > 0$ and $\rho_\star > 0$ such that*

$$\|\mathcal{A}_\Omega(M - M^\star)\|_{\mathrm{F}}^2 \ge \alpha_\star \min\{\|M - M^\star\|_{\mathrm{F}}^2, \rho_\star^2\} \tag{38}$$

for every $M \succeq 0$ with $\mathrm{rank}(M) \le r$ and $\|M - M^\star\|_{\mathrm{F}} \le D$. Then, for every $0 < \rho \le D$,

$$\mu_{\mathrm{id},D}(\rho) \ge \alpha_\star \min\left\{1, \frac{\rho_\star^2}{D^2}\right\}. \tag{39}$$

Remark C.3 (Interpretation of the stronger assumption)*.* Assumption (38) strengthens uniqueness on the bounded feasible class by requiring a quantitative observation gap: quadratic growth near $M^\star$, and a positive separation beyond $\rho_\star$ within the D-ball. This assumption is used only for (39); the recovery theorems use the bounded-modulus gap directly.

Lemma C.4 (Concentration of missing energy)**.** *Assume $\Omega \sim \mathrm{Ber}(p)$ entrywise. For every $\chi > 0$,*

$$\mathbb{P}\big(\big|\|\mathcal{A}_{\Omega^{\mathrm{c}}} M^\star\|_{\mathrm{F}}^2 - (1-p)\|M^\star\|_{\mathrm{F}}^2\big| > \chi\|M^\star\|_{\mathrm{F}}^2\big) \le 2\exp\left[-c\min\left\{\frac{\chi^2 s_4(M^\star)}{p(1-p)},\ \chi s_\infty(M^\star)\right\}\right]. \tag{40}$$

In particular, with the same probability,

$$\|\mathcal{A}_{\Omega^{\mathrm{c}}} M^\star\|_{\mathrm{F}}^2 \le (1 - p + \chi)\|M^\star\|_{\mathrm{F}}^2. \tag{41}$$

For the shell $\mathcal{H}(\rho)$, define the regularity parameters

$$s_{4,\rho} := \inf_{H \in \mathcal{H}(\rho)\setminus\{0\}} \frac{\|H\|_{\mathrm{F}}^4}{\|H\|_4^4}, \qquad s_{\infty,\rho} := \inf_{H \in \mathcal{H}(\rho)\setminus\{0\}} \frac{\|H\|_{\mathrm{F}}^2}{\|H\|_\infty^2}. \tag{42}$$

Theorem C.5 (Uniform Bernstein lower bound on the fixed shell)**.** *For every $\chi \in (0, p)$,*

$$\mathbb{P}\big(\|\mathcal{A}_\Omega H\|_{\mathrm{F}}^2 \ge (p - \chi)\|H\|_{\mathrm{F}}^2 \quad \textit{for all } H \in \mathcal{H}(\rho)\big) \ge 1 - \delta_{\mathrm{shell}}(\chi, \rho), \tag{43}$$

where

$$\delta_{\mathrm{shell}}(\chi, \rho) := 2\exp\left[(4n+2)r\log\left(\frac{36}{\chi}\right) - c\min\left\{\frac{\chi^2 s_{4,\rho}}{p(1-p)},\ \chi s_{\infty,\rho}\right\}\right]. \tag{44}$$

Proposition C.6 (GNMR radius forces shell regularity)**.** *Assume $M^\star$ is rank r and μ_0-incoherent. Let*

$$\rho_{\mathrm{reg}} := \min\left\{\frac{\sigma_r(M^\star)}{2},\ \frac{c_\mu\sqrt{\mu_0 r/n}}{1 + c_\mu\sqrt{\mu_0 r/n}}\,\sigma_r(M^\star)\right\}, \qquad c_\mu := \frac{\sqrt{3/2} - 1}{\sqrt{2}}.$$

If $\rho \le \rho_{\mathrm{reg}}$, then every nonzero $H = M - M^\star \in \mathcal{H}(\rho)$ satisfies

$$s_\infty(H) \ge \frac{n}{16\mu_0 r}, \qquad s_4(H) \ge \frac{n}{16\mu_0 r}.$$

Corollary C.7 (Explicit shell Bernstein bound under $\rho \le \rho_{\mathrm{GNMR}}$)**.** *Assume the hypotheses of Proposition C.6. Then, for every $\chi \in (0, p)$,*

$$\mathbb{P}\big(\|\mathcal{A}_\Omega H\|_{\mathrm{F}}^2 \ge (p - \chi)\|H\|_{\mathrm{F}}^2 \quad \textit{for all } H \in \mathcal{H}(\rho)\big) \ge 1 - \delta_{\mathrm{shell}}^{\mathrm{GNMR}}(\chi), \tag{45}$$

where

$$\delta_{\mathrm{shell}}^{\mathrm{GNMR}}(\chi) := 2\exp\left[(4n+2)r\log\left(\frac{36}{\chi}\right) - c\min\left\{\frac{\chi^2 n}{16p(1-p)\mu_0 r},\ \frac{\chi n}{16\mu_0 r}\right\}\right]. \tag{46}$$

Lemma C.8 (GNMR local RIP on the shell)**.** *Assume $M^\star \in \mathcal{M}(n, n, r, \mu_0, \kappa)$ and the sample complexity condition* (14)*. If $\rho \le \rho_{\mathrm{GNMR}}$, then with probability at least $1 - 3/n^3$,*

$$\|\mathcal{A}_\Omega H\|_{\mathrm{F}}^2 \ge p(1-\eta)^2\|H\|_{\mathrm{F}}^2 \qquad \textit{for all } H \in \mathcal{H}(\rho). \tag{47}$$

C.2 Missing Proofs

Proof of Lemma 2.1. By construction,

$$\|\mathcal{A}_{\Omega,\varepsilon}(M)\|_2^2 = \varepsilon^2\|M\|_{\mathrm{F}}^2 + (1-\varepsilon^2)\|\mathcal{A}_\Omega M\|_{\mathrm{F}}^2.$$

Since $0 \le \|\mathcal{A}_\Omega M\|_{\mathrm{F}}^2 \le \|M\|_{\mathrm{F}}^2$, the deterministic bounds follow. Under Bernoulli sampling, $\mathbb{E}\|\mathcal{A}_\Omega M\|_{\mathrm{F}}^2 = p\|M\|_{\mathrm{F}}^2$, proving the expectation identity. For fixed $M \neq 0$, let $Z = \sum_{ij}(\delta_{ij}-p)M_{ij}^2$, where the δ_{ij} are independent Bernoulli variables. Its summands are bounded in magnitude by $B = \|M\|_\infty^2$, and its variance is at most $p(1-p)B\|M\|_{\mathrm{F}}^2$. One-sided Bernstein therefore gives, with probability at least $1-\delta$,

$$Z \ge -\sqrt{2p(1-p)B\|M\|_{\mathrm{F}}^2 u} - \frac{2Bu}{3}.$$

Dividing by $\|M\|_{\mathrm{F}}^2$, combining with $\|\mathcal{A}_\Omega M\|_{\mathrm{F}}^2 \ge 0$, and substituting into the first identity proves (6). □

Proof of Lemma 3.1. For fixed $0 < \rho \le D$, the annulus in (8) is compact and the denominator is at least $\rho^2 > 0$. Uniqueness within the D-ball makes its continuous numerator strictly positive, so the ratio attains a positive minimum. Conversely, any distinct feasible completion at distance $d \in (0, D]$ has zero numerator and gives $\mu_{\mathrm{id},D}(d) = 0$. Global unique completability implies uniqueness within this ball. □

Proof of Lemma C.1. Since $M^\star$ is feasible for (5), global optimality of $\widetilde{M}$ yields

$$\frac{1}{2}\|\mathcal{A}_\Omega(\widetilde{M} - M^\star)\|_{\mathrm{F}}^2 + \frac{\varepsilon^2}{2}\|\mathcal{A}_{\Omega^c}\widetilde{M}\|_{\mathrm{F}}^2 \le \frac{1}{2}\|\mathcal{A}_\Omega(M^\star - M^\star)\|_{\mathrm{F}}^2 + \frac{\varepsilon^2}{2}\|\mathcal{A}_{\Omega^c}M^\star\|_{\mathrm{F}}^2.$$

Since $\widetilde{M} = \Delta + M^\star$, this is exactly (34). Dropping the second term proves (35); dropping the first proves (36).

Finally, mask orthogonality gives $\|\Delta\|_{\mathrm{F}}^2 = \|\mathcal{A}_\Omega\Delta\|_{\mathrm{F}}^2 + \|\mathcal{A}_{\Omega^c}\Delta\|_{\mathrm{F}}^2$. By the triangle inequality and (36), $\|\mathcal{A}_{\Omega^c}\Delta\|_{\mathrm{F}} \le 2\|\mathcal{A}_{\Omega^c}M^\star\|_{\mathrm{F}}$, proving (37). □

Proof of Lemma C.2. For $0 < \rho_1 \le \rho_2 \le D$, the annulus with lower radius ρ_2 is a subset of that with lower radius ρ_1. Taking the infimum of the same ratio over the smaller set gives $\mu_{\mathrm{id},D}(\rho_1) \le \mu_{\mathrm{id},D}(\rho_2)$. For any competitor in the annulus, write $d = \|M - M^\star\|_{\mathrm{F}}$. Assumption (38) gives a ratio at least $\alpha_\star$ when $d \le \rho_\star$, and at least $\alpha_\star\rho_\star^2/d^2 \ge \alpha_\star\rho_\star^2/D^2$ otherwise. Taking the infimum proves (39), including the case $\rho_\star \ge D$. □

Proof of Lemma C.4. Write

$$\|\mathcal{A}_{\Omega^c}M^\star\|_{\mathrm{F}}^2 = \sum_{i,j}(1-\delta_{ij})M^{\star 2}_{ij}, \qquad \delta_{ij} \sim \mathrm{Ber}(p) \text{ independently.}$$

Define

$$X_{ij} := \big((1-\delta_{ij}) - (1-p)\big)M^{\star 2}_{ij}.$$

Then $\mathbb{E}[X_{ij}] = 0$ and

$$\sum_{i,j} X_{ij} = \|\mathcal{A}_{\Omega^c}M^\star\|_{\mathrm{F}}^2 - (1-p)\|M^\star\|_{\mathrm{F}}^2.$$

Moreover,

$$|X_{ij}| \le \|M^\star\|_\infty^2, \qquad \sum_{i,j}\mathrm{Var}(X_{ij}) = p(1-p)\sum_{i,j}M^{\star 4}_{ij} = p(1-p)\|M^\star\|_4^4.$$

Bernstein's inequality therefore gives

$$\mathbb{P}\left(\left|\sum_{i,j}X_{ij}\right| > u\right) \le 2\exp\left[-\frac{u^2/2}{p(1-p)\|M^\star\|_4^4 + \|M^\star\|_\infty^2 u/3}\right].$$

Set $u = \chi\|M^\star\|_{\mathrm{F}}^2$ and use

$$\frac{\|M^\star\|_{\mathrm{F}}^4}{\|M^\star\|_4^4} = s_4(M^\star), \qquad \frac{\|M^\star\|_{\mathrm{F}}^2}{\|M^\star\|_\infty^2} = s_\infty(M^\star),$$

to obtain (40). The upper bound (41) follows immediately. □

Proof of Lemma 3.2. Suppose for contradiction that $\|\Delta\|_{\mathrm{F}} \geq \rho$. By (37), $\|\Delta\|_{\mathrm{F}} \leq D$, so $\widetilde{M}$ lies in the defining annulus. By definition of $\mu_{\mathrm{id},D}(\rho)$,

$$\|\mathcal{A}_\Omega \Delta\|_{\mathrm{F}}^2 \geq \mu_{\mathrm{id},D}(\rho)\|\Delta\|_{\mathrm{F}}^2 \geq \mu_{\mathrm{id},D}(\rho)\rho^2.$$

On the other hand, (35) from Lemma C.1 gives

$$\|\mathcal{A}_\Omega \Delta\|_{\mathrm{F}}^2 \leq \varepsilon^2 \|\mathcal{A}_{\Omega^c} M^\star\|_{\mathrm{F}}^2,$$

contradicting (9). Hence $\|\Delta\|_{\mathrm{F}} < \rho$. □

Proof of Theorem C.5. Let

$$\mathcal{U}(\rho) := \left\{ \frac{H}{\|H\|_{\mathrm{F}}} : \ H \in \mathcal{H}(\rho) \setminus \{0\} \right\}.$$

Since every $H \in \mathcal{H}(\rho)$ is a difference of two rank-$\leq r$ matrices, one has $\mathrm{rank}(H) \leq 2r$. Therefore $\mathcal{U}(\rho)$ is contained in the unit Frobenius sphere of rank-$\leq 2r$ matrices. A standard covering argument (Candes and Plan, 2010) yields a $(\chi/4)$-net $\mathcal{N} \subseteq \mathcal{U}(\rho)$ with cardinality

$$|\mathcal{N}| \leq \left(\frac{36}{\chi}\right)^{(4n+2)r}.$$

Fix $G \in \mathcal{N}$. Since $\Omega \sim \mathrm{Ber}(p)$ entrywise and G is deterministic, the same scalar Bernstein argument as in Lemma C.4 gives

$$\mathbb{P}\big(\|\mathcal{A}_\Omega G\|_{\mathrm{F}}^2 < (p - \chi/2)\|G\|_{\mathrm{F}}^2\big) \leq 2\exp\left[-c\min\left\{\frac{\chi^2 s_4(G)}{p(1-p)},\ \chi s_\infty(G)\right\}\right].$$

Since $G \in \mathcal{U}(\rho)$, we have $s_4(G) \geq s_{4,\rho}$ and $s_\infty(G) \geq s_{\infty,\rho}$. Union bound over $\mathcal{N}$ yields

$$\mathbb{P}\big(\|\mathcal{A}_\Omega G\|_{\mathrm{F}}^2 \geq (p - \chi/2)\|G\|_{\mathrm{F}}^2 \text{ for all } G \in \mathcal{N}\big) \geq 1 - \delta_{\mathrm{shell}}(\chi, \rho).$$

Now let $H \in \mathcal{U}(\rho)$ and choose $G \in \mathcal{N}$ with $\|H - G\|_{\mathrm{F}} \leq \chi/4$. Then, on the event that

$$\|\mathcal{A}_\Omega G\|_{\mathrm{F}}^2 \geq (p - \chi/2)\|G\|_{\mathrm{F}}^2 \qquad \text{for all } G \in \mathcal{N},$$

we compare H to its nearby net point G. Since $H, G \in \mathcal{U}(\rho)$, we have

$$\|H\|_{\mathrm{F}} = \|G\|_{\mathrm{F}} = 1.$$

Also, for every matrix Z,

$$\|\mathcal{A}_\Omega Z\|_{\mathrm{F}} \leq \|Z\|_{\mathrm{F}},$$

because $\mathcal{A}_\Omega$ only keeps a subset of the entries of Z. Using

$$\big|\|X\|_F^2 - \|Y\|_F^2\big| = \big|\|X\|_F - \|Y\|_F\big|\big(\|X\|_F + \|Y\|_F\big) \leq \|X - Y\|_F\big(\|X\|_F + \|Y\|_F\big),$$

with $X = \mathcal{A}_\Omega H$ and $Y = \mathcal{A}_\Omega G$, we obtain

$$\begin{aligned}
\big|\|\mathcal{A}_\Omega H\|_{\mathrm{F}}^2 - \|\mathcal{A}_\Omega G\|_{\mathrm{F}}^2\big| &\leq \|\mathcal{A}_\Omega(H - G)\|_{\mathrm{F}}\big(\|\mathcal{A}_\Omega H\|_{\mathrm{F}} + \|\mathcal{A}_\Omega G\|_{\mathrm{F}}\big) \\
&\leq \|H - G\|_{\mathrm{F}}\big(\|H\|_{\mathrm{F}} + \|G\|_{\mathrm{F}}\big) \\
&\leq \frac{\chi}{4} \cdot 2 = \frac{\chi}{2}.
\end{aligned}$$

Hence, still on the net event,

$$\|\mathcal{A}_\Omega H\|_{\mathrm{F}}^2 \geq \|\mathcal{A}_\Omega G\|_{\mathrm{F}}^2 - \frac{\chi}{2} \geq \left(p - \frac{\chi}{2}\right)\|G\|_{\mathrm{F}}^2 - \frac{\chi}{2} = p - \chi.$$

Since $H \in \mathcal{U}(\rho)$ was arbitrary, this shows that

$$\|\mathcal{A}_\Omega H\|_{\mathrm{F}}^2 \geq p - \chi \qquad \text{for all } H \in \mathcal{U}(\rho)$$

with probability at least $1 - \delta_{\mathrm{shell}}(\chi, \rho)$. Scaling back from $\mathcal{U}(\rho)$ to $\mathcal{H}(\rho)$ proves (43). □

Proof of Proposition C.6. Let $0 \neq H = M - M^\star \in \mathcal{H}(\rho)$, put $h = \|H\|_{\mathrm{F}}$, $a = \mu_0 r/n$, and $\sigma = \sigma_r(M^\star)$. We show $\|H\|_\infty \le 4\sqrt{a}\, h$. If $a \ge 1/2$, this follows directly from $\|H\|_\infty \le h$. Otherwise $\mu_0 \ge 1$ implies $n > 2r$. Weyl's inequality and $h \le \rho \le \sigma/2$ imply $\operatorname{rank}(M) = r$. Write $M^\star = U_\star \Lambda_\star U_\star^T$ and $M = U\Lambda U^T$. The identity

$$(I - U_\star U_\star^T) U\Lambda = (I - U_\star U_\star^T) HU$$

gives

$$\|\sin\Theta(U, U_\star)\|_2 \le \frac{\rho}{\sigma - \rho} \le c_\mu \sqrt{a}.$$

Choose principal-vector bases, so that $U = U_\star C + U_\perp S$, where C, S are the diagonal cosine and sine matrices and $[U_\star\ U_\perp]$ has orthonormal columns. Such a completion is possible since $n > 2r$. These rotations preserve row norms and conjugate $\Lambda_\star, \Lambda$. Then

$$H = [U_\star\ U_\perp] \begin{bmatrix} A & B \\ B^T & D \end{bmatrix} [U_\star\ U_\perp]^T, \quad A = C\Lambda C - \Lambda_\star, \quad B = C\Lambda S, \quad D = S\Lambda S.$$

Orthogonality gives $h^2 = \|A\|_{\mathrm{F}}^2 + 2\|B\|_{\mathrm{F}}^2 + \|D\|_{\mathrm{F}}^2$, so $\|A\|_{\mathrm{op}}, \|B\|_{\mathrm{op}} \le h$. Since $\|S\|_{\mathrm{op}} \le c_\mu\sqrt{a} < 1/2$, C is invertible and $\|C^{-1}\|_{\mathrm{op}} \le 2$. The exact identity $D = SC^{-1}B$ implies $\|D\|_{\mathrm{op}} \le 2c_\mu\sqrt{a}\, h$. Using $\|e_i^T U_\star\|_2 \le \sqrt{a}$ and $\|e_i^T U_\perp\|_2 \le 1$ separately on each block gives

$$|H_{ij}| \le ah + 2\sqrt{a}\, h + 2c_\mu\sqrt{a}\, h \le (3 + 2c_\mu)\sqrt{a}\, h \le 4\sqrt{a}\, h.$$

Thus $s_\infty(H) \ge 1/(16a) = n/(16\mu_0 r)$. Finally, $\|H\|_4^4 \le \|H\|_\infty^2 \|H\|_{\mathrm{F}}^2$ gives $s_4(H) \ge s_\infty(H)$. $\square$

Proof to Lemma C.8. Recall the GNMR local-RIP neighborhood defined in Zilber and Nadler (2022a):

$$\mathcal{N}_{\mathrm{GNMR}}(\eta) := \Big\{ M \in \mathbb{R}^{n\times n} : \operatorname{rank}(M) = r,\ M \text{ is } \tfrac{3}{2}\mu_0\text{-incoherent},\ \|M - M^\star\|_{\mathrm{F}} \le \tfrac{\eta\sigma_r^\star}{c_e} \Big\},$$

where $\sigma_r^\star := \sigma_r(M^\star)$ and $c_e > 0$ is the absolute constant from the GNMR theorem. By the GNMR local RIP result, under the sample complexity condition (14), there exists an event $\mathcal{E}_{\mathrm{GNMR}}$ with

$$\mathbb{P}(\mathcal{E}_{\mathrm{GNMR}}) \ge 1 - \frac{3}{n^3}$$

such that on $\mathcal{E}_{\mathrm{GNMR}}$,

$$\|\mathcal{A}_\Omega(M - M^\star)\|_{\mathrm{F}}^2 \ge p(1-\eta)^2 \|M - M^\star\|_{\mathrm{F}}^2 \qquad \text{for all } M \in \mathcal{N}_{\mathrm{GNMR}}(\eta). \tag{48}$$

We will show that if $\rho \le \rho_{\mathrm{GNMR}}$, then every shell element $H \in \mathcal{H}(\rho)$ satisfies $M := M^\star + H \in \mathcal{N}_{\mathrm{GNMR}}(\eta)$, and then (48) will imply the claim.

Let $H \in \mathcal{H}(\rho)$ and define

$$M := M^\star + H.$$

By definition of the shell,

$$\|M - M^\star\|_{\mathrm{F}} = \|H\|_{\mathrm{F}} \le \rho.$$

Since $\rho \le \rho_{\mathrm{GNMR}} \le \eta\sigma_r^\star/c_e$, it follows that

$$\|M - M^\star\|_{\mathrm{F}} \le \frac{\eta\sigma_r^\star}{c_e}. \tag{49}$$

Because $H \in \mathcal{H}(\rho)$, the associated matrix $M = M^\star + H$ belongs to the admissible model class, so in particular

$$\operatorname{rank}(M) \le r.$$

Also, since $M^\star$ has rank r, Weyl's inequality yields

$$\sigma_r(M) \ge \sigma_r(M^\star) - \|M - M^\star\|_{\mathrm{op}} \ge \sigma_r^\star - \|M - M^\star\|_{\mathrm{F}} \ge \sigma_r^\star - \rho.$$

Using $\rho \le \rho_{\mathrm{GNMR}} \le \sigma_r^\star/2$, we obtain

$$\sigma_r(M) \ge \frac{\sigma_r^\star}{2} > 0.$$

Hence $\operatorname{rank}(M) \geq r$. Combined with $\operatorname{rank}(M) \leq r$, this gives

$$\operatorname{rank}(M) = r. \tag{50}$$

Write the eigendecompositions

$$M^\star = U_\star \Lambda_\star U_\star^\top, \qquad M = U \Lambda U^\top,$$

where $U_\star, U \in \mathbb{R}^{n \times r}$ have orthonormal columns. Since $M^\star \in \mathcal{M}(n, n, r, \mu_0, \kappa)$, we know that

$$\max_{1 \leq i \leq n} \|e_i^\top U_\star\|_2^2 \leq \frac{\mu_0 r}{n}.$$

By Davis–Kahan,

$$\|\sin\Theta(U, U_\star)\|_{\mathrm{op}} \leq \frac{\|M - M^\star\|_{\mathrm{op}}}{\sigma_r^\star - \rho} \leq \frac{\|M - M^\star\|_{\mathrm{F}}}{\sigma_r^\star - \rho} \leq \frac{\rho}{\sigma_r^\star - \rho}.$$

Therefore there exists an orthogonal matrix $Q \in \mathbb{R}^{r \times r}$ such that

$$\|U - U_\star Q\|_{\mathrm{op}} \leq \sqrt{2}\, \|\sin\Theta(U, U_\star)\|_{\mathrm{op}} \leq \sqrt{2}\, \frac{\rho}{\sigma_r^\star - \rho}.$$

Fix any row index i. Then

$$\|e_i^\top U\|_2 \leq \|e_i^\top U_\star Q\|_2 + \|e_i^\top (U - U_\star Q)\|_2 \leq \sqrt{\frac{\mu_0 r}{n}} + \sqrt{2}\, \frac{\rho}{\sigma_r^\star - \rho}.$$

Now choose the constant hidden in ρ_{GNMR} so that

$$\rho_{\mathrm{GNMR}} \leq \frac{c_\mu \sqrt{\mu_0 r / n}}{1 + c_\mu \sqrt{\mu_0 r / n}} \sigma_r^\star, \qquad c_\mu \leq \frac{\sqrt{3/2} - 1}{\sqrt{2}}.$$

Then, since $\rho \leq \rho_{\mathrm{GNMR}}$,

$$\sqrt{2}\, \frac{\rho}{\sigma_r^\star - \rho} \leq \Big(\sqrt{\frac{3}{2}} - 1\Big) \sqrt{\frac{\mu_0 r}{n}}.$$

Substituting into the previous row bound gives

$$\|e_i^\top U\|_2 \leq \sqrt{\frac{\mu_0 r}{n}} + \Big(\sqrt{\frac{3}{2}} - 1\Big) \sqrt{\frac{\mu_0 r}{n}} = \sqrt{\frac{3\mu_0 r}{2n}}.$$

Since this holds for every i,

$$\max_{1 \leq i \leq n} \|e_i^\top U\|_2^2 \leq \frac{3\mu_0 r}{2n}, \tag{51}$$

so M is $(3/2)\mu_0$-incoherent. From (49), (50), and (51), we conclude that

$$M = M^\star + H \in \mathcal{N}_{\mathrm{GNMR}}(\eta) \qquad \text{for every } H \in \mathcal{H}(\rho).$$

Therefore, on the event $\mathcal{E}_{\mathrm{GNMR}}$, applying (48) to $M = M^\star + H$ yields

$$\|\mathcal{A}_\Omega H\|_{\mathrm{F}}^2 = \|\mathcal{A}_\Omega(M - M^\star)\|_{\mathrm{F}}^2 \geq p(1-\eta)^2 \|M - M^\star\|_{\mathrm{F}}^2 = p(1-\eta)^2 \|H\|_{\mathrm{F}}^2$$

for all $H \in \mathcal{H}(\rho)$. This is exactly (47). □

Proof of Theorem 3.3. Let

$$\mathcal{E}_{\mathrm{miss}}(\chi) := \{\|\mathcal{A}_{\Omega^c} M^\star\|_{\mathrm{F}}^2 \leq (1 - p + \chi)\|M^\star\|_{\mathrm{F}}^2\}$$

and

$$\mathcal{E}_{\mathrm{shell}}(\chi) := \left\{\|\mathcal{A}_\Omega H\|_{\mathrm{F}}^2 \geq (p - \chi)\|H\|_{\mathrm{F}}^2 \text{ for all } H \in \mathcal{H}(\rho)\right\}.$$

By Lemma C.4,

$$\mathbb{P}(\mathcal{E}_{\mathrm{miss}}(\chi)) \geq 1 - 2\exp\left[-c \min\left\{\frac{\chi^2 s_4(M^\star)}{p(1-p)},\ \chi s_\infty(M^\star)\right\}\right].$$

By Theorem C.5,

$$\mathbb{P}(\mathcal{E}_{\mathrm{shell}}(\chi)) \geq 1 - \delta_{\mathrm{shell}}(\chi, \rho).$$

On $\mathcal{E}_{\mathrm{miss}}(\chi)$, assumption (11) implies

$$\mu_{\mathrm{id},D}(\rho)\rho^2 > \varepsilon^2 \|\mathcal{A}_{\Omega^c} M^\star\|_{\mathrm{F}}^2.$$

Hence Lemma 3.2 gives $\Delta \in \mathcal{H}(\rho)$. On the event $\mathcal{E}_{\mathrm{shell}}(\chi)$,

$$\|\mathcal{A}_\Omega \Delta\|_{\mathrm{F}}^2 \geq (p - \chi)\|\Delta\|_{\mathrm{F}}^2.$$

Combining this with (35) from Lemma C.1 and the bound from $\mathcal{E}_{\mathrm{miss}}(\chi)$ gives

$$(p - \chi)\|\Delta\|_{\mathrm{F}}^2 \leq \varepsilon^2 \|\mathcal{A}_{\Omega^c} M^\star\|_{\mathrm{F}}^2 \leq \varepsilon^2 (1 - p + \chi)\|M^\star\|_{\mathrm{F}}^2,$$

which proves (12). The probability bound follows from a union bound. □

Proof of Theorem 3.4. Let $\mathcal{E}_{\mathrm{miss}}(\chi)$ be as in the previous proof. By Lemma C.4,

$$\mathbb{P}(\mathcal{E}_{\mathrm{miss}}(\chi)) \geq 1 - 2\exp\left[-c\min\left\{\frac{\chi^2 s_4(M^\star)}{p(1-p)},\ \chi s_\infty(M^\star)\right\}\right].$$

On this event, assumption (iv) implies

$$\mu_{\mathrm{id},D}(\rho)\rho^2 > \varepsilon^2 \|\mathcal{A}_{\Omega^c} M^\star\|_{\mathrm{F}}^2,$$

so Lemma 3.2 gives $\Delta \in \mathcal{H}(\rho)$. Since $\rho \leq \rho_{\mathrm{GNMR}}$, Corollary C.8 yields, with probability at least $1 - 3/n^3$,

$$\|\mathcal{A}_\Omega \Delta\|_{\mathrm{F}}^2 \geq p(1 - \eta)^2 \|\Delta\|_{\mathrm{F}}^2.$$

Combining this with (35) from Lemma C.1 and the event $\mathcal{E}_{\mathrm{miss}}(\chi)$ gives

$$p(1 - \eta)^2 \|\Delta\|_{\mathrm{F}}^2 \leq \varepsilon^2 \|\mathcal{A}_{\Omega^c} M^\star\|_{\mathrm{F}}^2 \leq \varepsilon^2 (1 - p + \chi)\|M^\star\|_{\mathrm{F}}^2,$$

which is (16). The probability bound follows by intersecting the GNMR event with $\mathcal{E}_{\mathrm{miss}}(\chi)$. □

Proof of Theorem 4.1. First of all, we hope to decompose the Hessian of (1) at a second order point $\hat{X} \in \mathbb{R}^{n \times r}$. Classic matrix sensing literatures like Ha et al. (2020); Zhang et al. (2021); Li et al. (2019) give that the second-order critical condition of (1) are given as

$$\nabla f(\hat{X}\hat{X}^\top)\hat{X} = 0, \tag{52}$$

$$2\langle \nabla f(\hat{X}\hat{X}^\top), UU^\top \rangle + [\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top + U\hat{X}^\top, \hat{X}U^\top + U\hat{X}^\top) \geq 0 \quad \forall U \in \mathbb{R}^{n \times r} \tag{53}$$

with (52) being the first order critical condition. Moreover, since the sensing matrices $\{A_i\}_{i \in [m]}$ can be assumed be to symmetric without loss of generality (Zhang et al., 2021), we have that

$$[\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top + U\hat{X}^\top, \hat{X}U^\top + U\hat{X}^\top) = 4[\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top, \hat{X}U^\top).$$

We then could decompose LHS of (53) as $2C_1 + 4C_2$ where

$$C_1 \coloneqq \langle \nabla f(\hat{X}\hat{X}^\top), UU^\top \rangle, \quad C_2 \coloneqq [\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top, \hat{X}U^\top)$$

Given the assumption that (1) obeys some RSS condition, it is possible to upper-bound C_2 by observing

$$[\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top + U\hat{X}^\top, \hat{X}U^\top + U\hat{X}^\top) \leq L_s \|\hat{X}U^\top + U\hat{X}^\top\|_F^2$$

Therefore, if want to somehow create an negative escape direction for $\hat{X}$, it is important that we find a U such that C_1 is negative and large in magnitude, and then amplify this term via tensor parametrization. To do so, we first do a more in-depth analysis of $\nabla f(\hat{X}\hat{X}^\top)$. As mentioned above, since $\nabla f(\cdot)$ can be assumed to be symmetric, one can select $u \in \mathbb{R}^n$ such that $u^\top \nabla f(\hat{x}\hat{x}^\top)u = \lambda_{\min}(\nabla f(\hat{x}\hat{x}^\top))$. Then via the definition of RSC we have

$$f(M^*) \geq f(\hat{X}\hat{X}^\top) + \langle \nabla f(\hat{X}\hat{X}^\top), M^* - \hat{X}\hat{X}^\top \rangle + \frac{\alpha_s}{2}\|\hat{X}\hat{X}^\top - M^*\|_F^2. \tag{54}$$

With $\hat{X}$ being a first-order point, according to (52)

$$\nabla f(\hat{X}\hat{X}^\top)\hat{X} = 0 \implies \langle \nabla f(\hat{X}\hat{X}^\top), \hat{X}\hat{X}^\top \rangle = 0$$

Therefore, if in (1) our b is corrupted as $\mathcal{A}(M^*) + \tilde{w}$, then plugging it back into (54) gives

$$\begin{aligned}
\langle \nabla f(\hat{X}\hat{X}^\top), M^* \rangle &\leq -\frac{\alpha_s}{2}\|\hat{x}\hat{x}^\top - M^*\|_F^2 + f(M^*) - f(XX^\top) \\
&\leq -\frac{\alpha_s}{2}\|\hat{x}\hat{x}^\top - M^*\|_F^2 + f(M^*) \\
&= -\frac{\alpha_s}{2}\|\hat{x}\hat{x}^\top - M^*\|_F^2 + \frac{\|\tilde{w}\|_2^2}{2}
\end{aligned} \tag{55}$$

where the second inequality follows from the fact that $f(\cdot) \geq 0$ in its entire domain and the last inequality follows from $f(M^*) = 1/2\|\mathcal{A}(M^*) - \mathcal{A}(M^*) - \tilde{w}\|_2^2 = \|\tilde{w}\|_2^2/2$. Furthermore, since both $\nabla f(\hat{X}\hat{X}^\top)$ and M^* are assumed to be positive semidefinite,

$$\langle \nabla f(\hat{X}\hat{X}^\top), M^* \rangle \geq \lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \operatorname{tr}(M^*)$$

which implies that

$$\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \leq \frac{-\alpha_s\|\hat{X}\hat{X}^\top - M^*\|_F^2 + \|\tilde{w}\|_2^2}{2\operatorname{tr}(M^*)} \tag{56}$$

Furthermore, (19) gives us

$$\|\hat{X}\hat{X}^\top - M^*\|_F^2 \geq \|\tilde{w}\|_2^2/\alpha_s$$

since $\frac{L_s}{\alpha_s}\lambda_r(\hat{X}\hat{X}^\top)\operatorname{tr}(M^*) \geq 0$ by definition. This means that

$$\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \leq \frac{-\alpha_s\|\hat{X}\hat{X}^\top - M^*\|_F^2 + \|\tilde{w}\|_2^2}{2\operatorname{tr}(M^*)} \leq 0 \tag{57}$$

Thus, with this result equipped, we can further find a U that makes C_1 small. In the most convenient manner, we first consider the eigenvector $u \in \mathbb{R}^n$ of $\nabla f(\hat{X}\hat{X}^\top)$ associated with $\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top))$. Additionally we consider $q \in \mathbb{R}^r$ to be the r-th singular value of $\hat{X}$, with

$$\|\hat{X}q\|_2 = \sigma_r(\hat{X}), \qquad \|q\|_2 = 1$$

Then choosing $U \in \mathbb{R}^{n\times r} = uq^\top$ leads to

$$C_1 = \langle \nabla f(\hat{X}\hat{X}^\top), UU^\top \rangle = \langle \nabla f(\hat{X}\hat{X}^\top), uu^\top \rangle = -G$$

where $G \coloneqq -\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \geq 0$. By recalling $\hat{X}^\top u = 0$ according to the first-order condition (52), we can further bound C_2 with this chocie of U as

$$\begin{aligned}
L_s\|\hat{X}U^\top + U\hat{X}^\top\|_F^2 &= L_s\|u(\hat{X}q)^\top + (\hat{X}q)u^\top\|_F^2 \\
&= 2L_s\|\hat{X}q\|_F^2 + 2L_s(q^\top(\hat{X}^\top u))^2 \\
&= 2L_s\lambda_r(\hat{X}\hat{X}^\top),
\end{aligned}$$

leading to

$$C_2 \leq \frac{1}{2}L_s\lambda_r(\hat{X}\hat{X}^\top)$$

Now, if we choose $\Delta = \operatorname{vec}(U)^{\otimes l}$ for the aforementioned $U \in \mathbb{R}^{n\times r}$, the LHS of (27b) can be expressed as:

$$\begin{aligned}
&2(\langle \mathbf{A}, \hat{X}\hat{X}^\top \rangle_{2,3}^\top \langle \mathbf{A}, uu^\top \rangle_{2,3})^l - 2\left((\langle \mathbf{A}, M^* \rangle_{2,3} + \tilde{w})^\top \langle \mathbf{A}, uu^\top \rangle_{2,3}\right)^l + 4(\|\langle \mathbf{A}, \hat{X}U^\top \rangle_{2,3}\|_2^2)^l \\
\leq& 2(\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)))^l + 4C_2^l \\
=& 2C_1^l + 4C_2^l
\end{aligned} \tag{58}$$

where the inequality follows from:

$$a^n - b^n \leq (a-b)^n, \quad \forall b \geq a \geq 0$$

Here, since $a - b = C_1 \leq 0$, the above inequality can be used. As a result,

$$\text{LHS of (27b)} \leq \underbrace{-2G^l}_{\text{Part 1}} + \underbrace{\frac{2}{2^{l-1}} L_s^l \lambda_r(\hat{X}\hat{X}^\top)^l}_{\text{Part 2}}$$

We know since $G \geq 0$, Part 1 is always negative assuming l is odd, and Part 2 is always positive. Therefore, it suffices to find an order l such that

$$G^l > (1/2^{l-1}) L_s^l \lambda_r(\hat{X}\hat{X}^\top)^l \tag{59}$$

Conveniently, (57) says that

$$G \geq \frac{\alpha_s \|M^* - \hat{X}\hat{X}^\top\|_F^2 - \|\tilde{w}\|_2^2}{2\,\mathrm{tr}(M^*)}, \tag{60}$$

which can be used to derive sufficient condition for (59). Therefore, if

$$\left(\frac{\alpha_s \|M^* - \hat{X}\hat{X}^\top\|_F^2 - \|\tilde{w}\|_2^2}{2\,\mathrm{tr}(M^*)} \right)^l > (1/2^{l-1}) L_s^l \lambda_r(\hat{X}\hat{X}^\top)^l,$$

we can conclude that (59) holds, which implies that the LHS of (27b) is negative, directly proving that $\hat{X}^{\otimes l}$ is not an SOP anymore. Elementary manipulations of the above equation give that a sufficient condition is

$$\|M^* - \hat{X}\hat{X}^\top\|_F^2 - \|\tilde{w}\|_2^2/\alpha_s > 2^{1/l} \frac{L_s}{\alpha_s} \lambda_r(\hat{X}\hat{X}^\top)\,\mathrm{tr}(M^*) \tag{61}$$

We now consider (19), which means that

$$\lambda_r(\hat{X}\hat{X}^\top) \leq \frac{\alpha_s \|M^* - \hat{X}\hat{X}^\top\|_F^2 - \|\tilde{w}\|_2^2}{L_s\,\mathrm{tr}(M^*)} \tag{62}$$

Subsequently, define a constant γ such that

$$L_s \lambda_r(\hat{X}\hat{X}^\top) = \gamma \left[\frac{\alpha_s \|M^* - \hat{X}\hat{X}^\top\|_F^2 - \|\tilde{w}\|_2^2}{2\,\mathrm{tr}(M^*)} \right]$$

Then, (60) and (62) together imply that $1 \leq \gamma < 2$. Using this simplified notation, our sufficient condition (61) becomes

$$1 > \frac{\gamma}{2^{(l-1)/l}} \tag{63}$$

Given $1 \leq \gamma < 2$, there always exists a large enough l such that (63) holds, which proves that LHS of (27b) is negative, proving that $\mathrm{vec}(\hat{X})^{\otimes l}$ is a strict saddle, concluding the proof.

To derive a sufficient l, we simply acknowledge

$$\gamma = \frac{2L_s\,\mathrm{tr}(M^*)\lambda_r(\hat{X}\hat{X}^\top)}{\alpha_s \|M^* - \hat{X}\hat{X}^\top\|_F^2 - \|\tilde{w}\|_2^2} := 2\beta$$

and that $\beta \leq 1$ due to assumption (19). Therefore, for (63) to hold true, it is enough to have

$$2^{(l-1)/l} > 2\beta \implies \frac{l-1}{l} > \log_2(2\beta) \implies l > \frac{1}{1 - \log_2(2\beta)}$$

□

Proof of Theorem B.9. First of all, we hope to decompose the GD trajectory of (18) $\{\mathbf{w}_t\}_{t=0}^T$ as follows:

$$\mathbf{w}_{t+1} = \langle \mathbf{Z}_t, \mathbf{w}_0 \rangle - \mathbf{E}_t := \tilde{\mathbf{w}}_t - \mathbf{E}_t \tag{64}$$

where

$$\begin{aligned}
\mathbf{Z}_t &:= (\mathcal{I} + \eta \langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l} \rangle)^t, \quad \mathbf{A}_r = I_r \oslash_{2,3} A \\
\mathbf{E}_t &:= \sum_{i=1}^t (\mathcal{I} + \eta \langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l} \rangle)^{t-i} \hat{\mathbf{E}}_i \\
\hat{\mathbf{E}}_i &:= \eta \langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}(\mathbf{w}_{i-1}), \mathbf{P}(\mathbf{w}_{i-1}) \rangle_{2*[l]} \rangle, \mathbf{w}_{i-1} \rangle_{2*[l]} \\
(\mathbf{A}_r^l)^* \mathbf{A}^l &:= \langle (\mathbf{A}_r)^{\otimes l}, \mathbf{A}^{\otimes l} \rangle_{3,6,\ldots,3l} \in \mathbb{R}^{[nr \times nr \times n \times n] \circ l}
\end{aligned}$$

This can be proved via induction where

$$\begin{aligned}\mathbf{w}_1 &= \left(\mathcal{I} + \eta\langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l} - (\mathbf{A}^{\otimes l})^* \langle \mathbf{P}(\mathbf{w}_0), \mathbf{P}(\mathbf{w}_0)\rangle\right)\mathbf{w}_0 \\ &= (\mathcal{I} + \eta\langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l}\rangle)\mathbf{w}_0 - \eta\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}(\mathbf{w}_0), \mathbf{P}(\mathbf{w}_0)\rangle\rangle \mathbf{w}_0 \\ &= \langle \mathbf{Z}_1, \mathbf{w}_0\rangle - \mathbf{E}_1\end{aligned}$$

This serves as our base case, and the induction step can be proven as

$$\begin{aligned}\mathbf{w}_{t+1} &= \left(\mathcal{I} + \eta\langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l} - (\mathbf{A}^{\otimes l})^* \langle \mathbf{P}(\mathbf{w}_t), \mathbf{P}(\mathbf{w}_t)\rangle\right)\mathbf{w}_t \\ &= (\mathcal{I} + \eta\langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l}\rangle)\mathbf{w}_t - \eta\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}(\mathbf{w}_t), \mathbf{P}(\mathbf{w}_t)\rangle\rangle \mathbf{w}_t \\ &= (\mathcal{I} + \eta\langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l}\rangle)\mathbf{w}_t - \hat{\mathbf{E}}_{t+1} \\ &= (\mathcal{I} + \eta\langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l}\rangle)\left(\tilde{\mathbf{w}}_t - \sum_{i=1}^{t} (\mathcal{I} + \eta\langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l}\rangle)^{t-i}\hat{\mathbf{E}}_i\right) - \hat{\mathbf{E}}_{t+1} \\ &= \tilde{\mathbf{w}}_{t+1} - \sum_{i=1}^{t} (\mathcal{I} + \eta\langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l}\rangle)^{t+1-i}\hat{\mathbf{E}}_i - \hat{\mathbf{E}}_{t+1} \\ &= \tilde{\mathbf{w}}_{t+1} - \sum_{i=1}^{t+1} (\mathcal{I} + \eta\langle \mathbf{A}_r^{\otimes l}, \tilde{b}^{\otimes l}\rangle)^{t+1-i}\hat{\mathbf{E}}_i \\ &= \tilde{\mathbf{w}}_{t+1} - \mathbf{E}_t\end{aligned}$$

Therefore, we can then use a version of Lemma 13 and Lemma 2 in Ma et al. (2024) with $U = \langle \mathbf{A}_r^* \mathbf{A}, M^* \rangle$ replaced with $U = \langle \mathbf{A}_r, \tilde{b}\rangle_1$ to prove this theorem, following the steps in proof to Theorem 1 in Ma et al. (2024). □

D Spatial distribution of reconstruction error

Global metrics such as the Frobenius error $\|M^{\text{sol}} - M^\star\|_F$ summarize the quality of reconstruction in a single number, but they do not reveal how errors are distributed across the matrix. In such cases, it is natural to ask whether failure manifests as a few "bad" rows or columns, or whether the algorithm fits the observed entries well but extrapolates incorrectly on a subset of unobserved entries.

To investigate this question, we analyze the entrywise squared error $E_{ij}^2 = (M_{ij}^{\text{sol}} - M_{ij}^\star)^2$ and its spatial distribution over the matrix. We consider the structured instance of problem (24) and use our method described in Theorem 4.3 to obtain reconstructed solutions M^{sol}. We performed numerical studies on E, and have made some interesting observations. A striking feature of all unsuccessful runs is that the error on observed entries is essentially zero. The optimization procedure consistently recovers the observed data very accurately. The non-negligible reconstruction error is confined to unobserved entries, indicating that the algorithm converges to stationary points that are locally optimal with respect to the observed data, rather than diverging or producing a globally incoherent factorization.

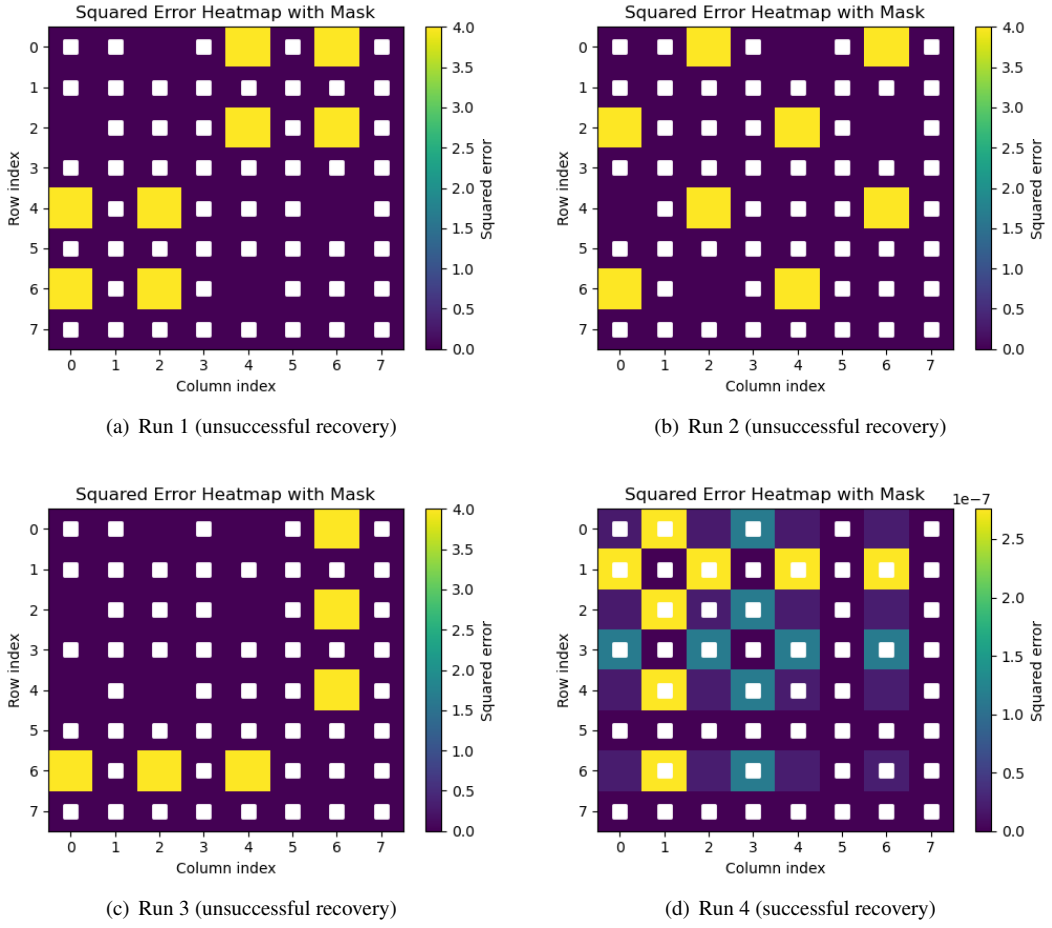
Table 3 showcases this phenomenon quantitatively. Across 10 independent runs, we report the mean-squared error on observed and unobserved entries, denoted MSE_{obs} and $\text{MSE}_{\text{unobs}}$, respectively. In 10 out of 10 runs, MSE_{obs} is on the order of 10^{-8} or even lower, while $\text{MSE}_{\text{unobs}}$ is solely responsible in determining whether the reconstruction was successful or not. This shows a gap of roughly eight orders of magnitude between the observed and unobserved regions. To take a closer look at the results, we selected 4 runs out of the 10 and plotted their entrywise reconstruction errors using heat-maps in Figure 3. We overlayed the observation mask Ω with white squares indicating the observed entries, while the background color encodes the squared reconstruction error.

From Figure 3 we observe that the error patterns on the unobserved entries differ substantially across the three unsuccessful runs. In each case, different subsets of the unobserved region carry the dominant error, and there is no single block or band that is systematically problematic across runs. This behavior is consistent with the presence of many spurious local minima in problem (24): different random initializations lead gradient descent to distinct low-rank completions that interpolate the observed entries but disagree on how to fill in the missing part. In contrast, the successful run

exhibits uniformly small squared error on both observed and unobserved entries, corresponding to convergence to the global minimum. Taken together, these experiments suggest that unsuccessful recoveries are primarily due to convergence to benign local minima that fit the data but differ in their extrapolation to unobserved entries, rather than to systematic failure in any specific region of the matrix.

	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9
$\mathrm{MSE}_{\mathrm{obs}}$	3.26e-8	4.89e-8	9.46e-9	9.79e-9	1.35e-8	1.67e-8	1.07e-8	9.40e-9	1.41e-8	1.50e-8
$\mathrm{MSE}_{\mathrm{unobs}}$	2.67e0	4.95e-8	2.67e0	2.67e0	8.23e-8	8.24e-8	2.67e0	2.67e0	2.67e0	2.67e0

Table 3: Mean-squared error on observed and unobserved entries for the structured instance in problem (24) with $n = 8$, $\ell = 3$, $\epsilon = 10^{-4}$, learning rate 0.02, and initialization scale 10^{-4}. Columns $s0$–$s9$ correspond to independent runs (`sol_0`–`sol_9`). In most runs, $\mathrm{MSE}_{\mathrm{obs}}$ is near machine precision while $\mathrm{MSE}_{\mathrm{unobs}}$ is either of order 1 (unsuccessful completions) or of order 10^{-8} (successful completions), highlighting that unsuccessful runs fit the observed entries but differ in their extrapolation to unobserved entries.



(a) Run 1 (unsuccessful recovery)

(b) Run 2 (unsuccessful recovery)

(c) Run 3 (unsuccessful recovery)

(d) Run 4 (successful recovery)

Figure 3: Spatial distribution of squared reconstruction error for the structured instance in problem (24) with $n = 8$, $\ell = 3$, $\epsilon = 10^{-4}$, learning rate 0.02, and initialization scale 10^{-4}. The background colormap shows $(M^{\mathrm{sol}}_{ij} - M^{\star}_{ij})^2$, and white squares denote observed entries $(i, j) \in \Omega$. In the three runs with higher global error (a)–(c), the reconstruction is essentially exact on observed entries and the error is concentrated on unobserved entries, with different spatial patterns across runs. In the successful run (d), the error is small on both observed and unobserved entries.

E Conditional algorithmic guarantees

We make explicit how the pointwise escape result can support a perturbed-gradient-descent (pGD) argument, and the additional conditions needed. Throughout this section, h^l has the unhalved normalization in (18). For ϵ-MC, $\alpha_s = \epsilon^2$, $L_s = 1$, and $\tilde{w} = -\epsilon \mathcal{A}_{\Omega^c} M^\star$; only its squared norm enters the following bounds.

Escape curvature. For a second-order point $\hat{X}$ covered by Theorem 4.1, set

$$E := \|M^\star - \hat{X}\hat{X}^\top\|_{\mathrm{F}}, \qquad s := \lambda_r(\hat{X}\hat{X}^\top), \qquad a := \frac{\alpha_s E^2 - \|\tilde{w}\|_2^2}{2\,\mathrm{tr}(M^\star)}.$$

Assume $\mathrm{tr}(M^\star) > 0$, $a > 0$, and odd l with $\beta < 2^{-1/l}$, where β is defined in (20); thus $L_s s = 2a\beta$. The proof of Theorem 4.1 chooses unit vectors u, q so that $U = uq^\top$ and $\Delta = \mathrm{vec}(U)^{\otimes l}$ have unit Frobenius norm, $C_1 = -G$ with $G \geq a$, and $C_2 \leq L_s s/2$. Its half-squared-residual calculation gives

$$2C_1^l + 4C_2^l \leq -2G^l + 2^{2-l}(L_s s)^l \leq -2a^l(1 - 2\beta^l) < 0. \tag{65}$$

Passing to the unhalved loss doubles the Hessian, so the negative bound in (65) remains a conservative upper bound for $\lambda_{\min}(\nabla^2 h^l(\hat{\mathbf{w}}))$ at $\hat{\mathbf{w}} = \mathrm{vec}(\hat{X})^{\otimes l}$.

From pointwise curvature to a region. Fix an ambient ball $K_R := \{\mathbf{w} : \|\mathbf{w}\|_{\mathrm{F}} \leq R\}$. Since h^l is polynomial, its gradient and Hessian have finite Lipschitz constants L_h and ρ_h on K_R. Let $\mathcal{C}_R \subseteq K_R$ contain the covered lifts above with uniform margins $a \geq \underline{a} > 0$ and $\beta \leq \overline{\beta} < 2^{-1/l}$, and set

$$\lambda := 2\underline{a}^l(1 - 2\overline{\beta}^l) > 0.$$

At any $\mathbf{w} \in K_R$ within Frobenius distance $\tau \leq \lambda/(2\rho_h)$ of $\mathbf{w}_1 \in \mathcal{C}_R$, Weyl's inequality gives

$$\lambda_{\min}(\nabla^2 h^l(\mathbf{w})) \leq -\lambda + \rho_h\|\mathbf{w} - \mathbf{w}_1\|_{\mathrm{F}} \leq -\lambda/2. \tag{66}$$

Let $\mathcal{P}$ be the set of rank-one lifts where (19) fails. This protected set is not identified with the global minimizers. To obtain strict-saddle alternatives on a region $\mathcal{D}_R \subseteq K_R$, one additionally needs the following coverage condition: every $\mathbf{w} \in \mathcal{D}_R$ with $\|\nabla h^l(\mathbf{w})\|_{\mathrm{F}} < \theta$ and $\mathrm{dist}(\mathbf{w}, \mathcal{P}) > \zeta$ lies in the tube in (66). Theorem 4.1 establishes curvature at covered lifts of second-order points; it does not establish this condition at arbitrary small-gradient tensors.

Corollary E.1 (Conditional pGD conclusion)**.** *Assume this coverage condition holds. Run the ambient-space pGD algorithm of Jin et al. (2017), with its prescribed perturbation distribution, step size, trigger, and waiting schedule. Assume its trajectory remains in $\mathcal{D}_R$ and all perturbed points and local segments required by its escape analysis remain in K_R. For an admissible tolerance*

$$0 < \eta_g < \min\{\theta, \lambda^2/(4\rho_h)\},$$

with probability at least $1 - \delta$, pGD reaches a point within distance ζ of $\mathcal{P}$ in $\widetilde{\mathcal{O}}(L_h\Delta_h/\eta_g^2)$ iterations, where $\Delta_h := h^l(\mathbf{w}_0) - \inf_{\mathbf{w}} h^l(\mathbf{w})$. The suppressed factors are polylogarithmic in dimension, inverse failure probability, and the landscape and accuracy parameters.

Proof. Under the stated smoothness and containment conditions, the pGD guarantee gives $\|\nabla h^l(\mathbf{w})\|_{\mathrm{F}} \leq \eta_g$ and $\lambda_{\min}(\nabla^2 h^l(\mathbf{w})) \geq -\sqrt{\rho_h \eta_g} > -\lambda/2$. The strict tolerance excludes both the large-gradient alternative and (66); coverage therefore gives the claimed distance bound. □

Theorem B.9 concerns plain GD under small-initialization and finite-horizon conditions. Its spectral κ-rank-one property does not itself establish the Frobenius tube bound above or preserve that tube under random pGD perturbations. Similarly, the additional $\mathcal{O}(\sqrt{r}\,\kappa^{1/(2l)})$ recovery term in Theorem B.10 applies under that theorem's separate hypotheses. Coverage and containment for the actual perturbed trajectory remain additional requirements; replacing ambient perturbations by rank-one-aware kicks requires a separate analysis. Consequently, Corollary E.1 supplies a conditional iteration statement, rather than a complete runtime and recovery theorem for GD followed by Tensor PCA and factor extraction.

F Direct optimization of the surrogate

We separate the effect of perturbation from lifting by pairing ordinary MC ($\epsilon = 0$) and ϵ-MC ($\epsilon = 5 \times 10^{-5}$) within each solver, using the same problems, masks, and initializations. The smaller controls use rank-one $n = 8$ instances, Bernoulli sampling with $p = 0.15$ or the structured family in (24), and 100 runs per solver/family. Success means that the returned estimate has Frobenius error at most 0.05 relative to M^*. For reference GNMR (Algorithm 3 of Zilber and Nadler (2022a), $\alpha = -1$), we evaluate its returned $UV^\top$ estimate.

Only 0.4% of all paired binary success outcomes change. The rank-five Bernoulli results at $n = 200$ in Table 2 use 10 matched runs per cell. Across these 120 pairs, only two outcomes change, both for Adam. Perturbation alone therefore does not systematically resolve the optimization difficulty in these tests. These controls are separate from the original lifted-method comparisons in Figure 1.

The deterministic constants at this perturbation are $\alpha_s = \epsilon^2 = 2.5 \times 10^{-9}$ and $L_s = 1$, giving $L_s/\alpha_s = 4 \times 10^8$ and normalized RIP constant $(1 - \epsilon^2)/(1 + \epsilon^2) \approx 0.999999995$. Thus positive RSC does not place the problem in the small-RIP regime of conventional matrix-sensing landscape guarantees. This is a conditioning diagnostic, not a causal explanation of every observed failure.

G Simulated lifting for the structured setting

The direct-solver controls illustrate a practical distinction between surrogate accuracy and optimization: a small perturbation can preserve an accurate target while GD still converges to a spurious stationary point. Tensor lifting addresses this obstruction by exposing negative curvature at covered lifts of bad factorized points, including points that admit no local descent direction in the original factor space. Explicitly optimizing the lifted variable, however, requires $\mathcal{O}((nr)^l)$ storage. This cost limits the dimensions that can be studied with the direct tensor implementation even when the escape mechanism itself is useful.

Simulated Oracle Direction (SOD) (Shen et al., 2026) uses the lifted construction to derive escape candidates that can be computed in the original factor space. The higher-order tensor serves as a mathematical oracle for constructing these candidates rather than as an optimization variable that must be stored. When GD stalls, SOD evaluates the candidates using the original objective, accepts an objective-decreasing move, and resumes GD from the accepted point. This permits a finite escape move even when a local factor-space search has stalled. The procedure can be repeated at subsequent stationary points; its derivation and algorithmic details are given by Shen et al. (2026).

This connection suggests a practical route for using the lifted landscape without its dominant storage cost. We examine that route on the structured family in (24), augmenting ordinary GD with SOD while keeping the same factor variable throughout. The study reaches $n = 50, 100, 200$ without forming the lifted tensor and compares the resulting trajectories with GD under matched update budgets.

Let $M^* = zz^\top$, where $z_i = 1$ for odd i and 0 otherwise. To examine SOD in its noiseless weighted-sensing setting, this study uses

$$f_n(x) = \frac{1}{2}\|\mathcal{A}_\Omega(xx^\top - M^*)\|_2^2 + \frac{\epsilon_n^2}{2}\|\mathcal{A}_{\Omega^c}(xx^\top - \mathbf{1}\mathbf{1}^\top)\|_2^2, \tag{67}$$

with the all-ones vector $\mathbf{1} \in \mathbb{R}^n$ and $\epsilon_n = 0.48/\sqrt{\lceil n/2 \rceil - 1}$. The complement target is fixed at one independently of the missing data; in this structured family it coincides with the missing ground-truth entries. Thus (67) uses the same observation pattern but a different complement target from the zero-filled surrogate (5). The scale 0.48 was selected from $\{0.18, 0.24, 0.30, 0.36, 0.42, 0.48\}$ in separate 100-run screening at dimensions $20, 25, 30, 35, 40, 45, 50$; the reported trials use fresh seeds.

For each dimension, GD and GD+SOD share 500 Gaussian initializations normalized to Euclidean norm 10^{-2}. The step size is $\eta_n = \min\{0.02, 1/n\}$, with a base cap of $3000 \max\{1, n/50\}$ updates and stationarity tolerance 10^{-8}. After the GD prefix, SOD is eligible when the gradient norm is at most 2×10^{-5}. We allow at most five accepted SOD events, each followed by at most $1200 \max\{1, n/50\}$ GD updates. Plain GD receives exactly the same number of GD updates as its paired GD+SOD run, while SOD adds the escape computations. Success means $\|xx^\top - M^*\|_F \le 0.05$.

Table 4: SOD success rates over 500 paired starts per dimension, with matched GD-update budgets.

n	ϵ_n	GD (%)	GD+SOD (%)
50	0.097980	13.2	51.4
100	0.068571	2.0	46.4
200	0.048242	0.0	52.2

Table 4 shows substantial improvements at all three dimensions, with GD+SOD retaining roughly 46%–52% success through $n = 200$. These favorable results illustrate the applicability of simulated lifting to the structured setting and motivate implementations that use the lifted escape mechanism without explicit tensor storage.

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 claims are backed by Theorems appearing later in the text.

   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: The Limitations paragraph discusses the perturbation tradeoff, storage cost, and scope of the guarantees; individual results state their assumptions.

   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: Assumptions are stated with each result; proofs and supporting arguments appear in the main text and Appendices B, C, and E.

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: Experimental settings are given in Section 5 and the appendices. Code is linked at `https://github.com/MOFA-LAB/mc_as_noisyms`.

   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: Code and instructions are linked in the experiments section.

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: Section 5 describes the main experiments. Appendices A, D, F, and G give ablations, error visualizations, direct-solver controls, and the SOD protocol.

   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: We explained in the experiments section the number of experiments repeated and other related info.

   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: Yes, the computer used was described in the footnote.

   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: We did not violate any code of ethics. The paper is primarily of theoretical nature.

   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: [N/A]

    Justification: This is a theoretical paper.

    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: This is a theoretical paper.

    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: Baseline methods are cited in Section 5; attribution and license information for the reused implementations are provided in the public code repository.

    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: No new assets are introduced.

    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: Our research does not involve 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 is a theoretical paper so no need for IRB approval.

    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: LLM was not used for any of the core contents. It was mostly used for format and language checking.

    Guidelines:

    - 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.