Noisy Low-Rank Matrix Completion via Transformed L_1 Regularization and its Theoretical Properties

Kun Zhao[1] **Jiayi Wang**[1] **Yifei Lou**[2]

[1]The University of Texas at Dallas [2]The University of North Carolina at Chapel Hill

Abstract

This paper focuses on recovering an underlying matrix from its noisy partial entries, a problem commonly known as matrix completion. We delve into the investigation of a non-convex regularization, referred to as transformed L_1 (TL1), which interpolates between the rank and the nuclear norm of matrices through a hyper-parameter $a \in (0, \infty)$. While some literature adopts such regularization for matrix completion, it primarily addresses scenarios with uniformly missing entries and focuses on algorithmic advances. To fill in the gap in the current literature, we provide a comprehensive statistical analysis for the estimator from a TL1-regularized recovery model under general sampling distribution. In particular, we show that when a is sufficiently large, the matrix recovered by the TL1-based model enjoys a convergence rate measured by the Frobenius norm, comparable to that of the model based on the nuclear norm, despite the challenges posed by the non-convexity of the TL1 regularization. When a is small enough, we show that the rank of the estimated matrix remains a constant order when the true matrix is exactly low-rank. A trade-off between controlling the error and the rank is established through different choices of tuning parameters. The appealing practical performance of TL1 regularization is demonstrated through a simulation study that encompasses various sampling mechanisms, as well as two real-world applications. Additionally, the role of the hyper-parameter a on the TL1-based model is explored via experiments to offer guidance in practical scenarios.

1 INTRODUCTION

In the era of big data, low-rank matrix completion has become an indispensable and prevalent resource in various fields, such as machine learning and statistics. It addresses the challenge of estimating the missing entries of a partially contaminated observed matrix, where the low-rank property plays a pivotal role in avoiding ill-posedness. This technique is widely used for tasks like recommendation systems (Kang et al., 2016; Gurini et al., 2018; Chen et al., 2021; Zhang et al., 2021), signal processing (Weng and Wang, 2012; Zhang and Zhang, 2020; Chen et al., 2023; Yuchi et al., 2023), image recovery (Changjun et al., 2012; Cao et al., 2014; Zheng et al., 2024), computer vision (Ji et al., 2010; Jia et al., 2022), seismic imaging (Kumlu, 2021; Popa et al., 2021, 2022), and data imputation (Chen et al., 2017, 2020a; Xu et al., 2023).

To recover a low-rank matrix, matrix factorization, which reformulates the problem into a non-convex optimization task, has been extensively studied. For instance, alternating minimization methods for low-rank matrix decomposition have been explored by Jain et al. (2013) and Gu et al. (2023). To address challenges with ill-conditioned low-rank matrices, Tong et al. (2021) proposed a scaled gradient descent approach, offering improved estimation efficiency. Additionally, Ahn and Suarez (2021) examined matrix factorization techniques through the lens of Riemannian geometry, while Chen et al. (2022) introduced a non-convex framework for matrix completion with linearly parameterized factors, further enriching the landscape of low-rank matrix recovery methods.

Despite these numerical advances, matrix factorization remains inherently limited. First, its non-convex nature means there are no general guarantees of finding a global solution, making these methods particularly vulnerable to issues such as ill-conditioning, sensitivity to initialization, and challenges in selecting the appropriate rank (Keshavan et al., 2010; Jain et al., 2013). Second, while low-rank matrix recovery can be formulated as a rank minimization problem, this approach is

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

unfortunately NP-hard (Natarajan, 1995). As a result, existing algorithms for its exact solution are impractical due to the infeasibly large demands on time and computational resources (Chistov and Grigor'Ev, 1984). To address these challenges, a popular alternative is the convex relaxation of rank, referred to as the nuclear norm, which is defined to be the sum of the singular values of the matrix. A large body of literature (Recht et al., 2010; Candes and Plan, 2010; Candès and Tao, 2010; Koltchinskii et al., 2011; Recht, 2011; Gross, 2011; Candes and Recht, 2012; Klopp, 2014; Klopp et al., 2015) considers the nuclear-norm regularization for low-rank matrix completion and provides theoretical guarantees for its effectiveness.

However, much of the existing work operates under the assumption of a uniform missing structure, where every entry in the matrix is assumed to be observed with equal probability (Candes and Plan, 2010; Candès and Tao, 2010; Koltchinskii et al., 2011; Candes and Recht, 2012; Hastie et al., 2015; Bhaskar, 2016; Cherapanamjeri et al., 2017; Bi et al., 2017; Chen et al., 2020b; Xia and Yuan, 2021; Farias et al., 2022). In reality, this sampling assumption is often impractical for real-world applications. For instance, regarding the well-known Netflix Prize (Bennett et al., 2007), the dataset can be represented by a low-rank matrix with users as rows, movies as columns, and ratings as entry values. Certain movies are rated more frequently than others, and some users rate more movies than others, leading to a non-uniform distribution of known entries. Consequently, there has been increasing attention on non-uniform sampling, with related works generally falling into two categories: missing at random (MAR) (Srebro and Salakhutdinov, 2010; Király et al., 2015; Chen et al., 2015; Cho et al., 2017) and missing not at random (MNAR) (Ma and Chen, 2019; Sportisse et al., 2020; Jin et al., 2022; Li et al., 2024). Since the case of MNAR is known to be particularly challenging due to the identification issue (Little and Rubin, 2019), we primarily focus on the MAR scenario following the work (MacDonald, 2002).

Under the MAR setting, Klopp (2014) applied the standard nuclear norm penalty to deal with general sampling distributions and provided theoretical insights on the estimation errors measured in the Frobenius norm. However, it is reported empirically that the nuclear norm overestimates the matrix's rank (Wang et al., 2021). Alternatively, the max-norm was first proposed by Srebro et al. (2004) for matrix completion under non-uniform sampling mechanisms. Later, Cai and Zhou (2016) proved its theoretical superiority over the nuclear norm for noisy matrix completion under a general sampling model. Sequentially, Fang et al. (2018) proposed a more flexible estimator, called the hybrid regularizer, which incorporates both max-norm and nuclear norm. They demonstrated the performance of this hybrid regularizer in comparison to the max-norm and nuclear norm under various settings. However, the hybrid approach only achieves a sub-optimal rate for recovering exact low-rank matrices under a uniform sampling scheme, compared to the nuclear norm, and it comes at the cost of a high computational burden due to the use of semidefinite programming (SDP) (Srebro et al., 2004).

Another promising approach that balances computational efficiency and recovery performance is the transformed L_1 (TL1) regularization, originating in sparse recovery (Zhang and Xin, 2018; Ma et al., 2019). When applied to a vector, TL1 can interpolate between the L_0 semi-norm and the L_1 norm with a hyper-parameter $a \in (0, \infty)$. Consequently, when applied to the vector composed of the singular values of a matrix, TL1 can approximate both the rank and the nuclear norm of the matrix (Zhang et al., 2017). This non-convex regularization offers a closer approximation to the rank function than the convex nuclear norm, better capturing the low-rank structure of the matrix and retaining the computational complexity of nuclear norm minimization. Zhang et al. (2017) provided the convergence analysis of the numerical algorithm under a uniform sampling scheme. However, to the best of our knowledge, TL1 regularization has been empirically explored for low-rank matrix completion under uniform sampling, while its theoretical guarantees and recovery performance under more general sampling schemes have yet to be thoroughly examined.

In this paper, we provide a comprehensive statistical analysis of the estimator derived from a TL1-regularized recovery model under a general sampling distribution. Particularly, we demonstrate that for a sufficiently large a, the estimator of the target matrix using TL1 regularization achieves the optimal convergence rate on the Frobenius norm error, comparable to that of the nuclear norm-based model. This theoretical guarantee aligns with the property of the TL1 function: as $a \to \infty$, the TL1 function approximates the nuclear norm more closely. When a is small enough, we establish a sub-optimal convergence rate for the estimation error and an upper bound on the rank of the estimator. Our results imply that by choosing a smaller value of a, the estimated matrix tends to have a lower rank, which can be advantageous in applications where a low-rank solution is preferred for reasons such as simplicity or interpretability. This phenomenon is in line with another property of the TL1 function: as $a \to 0$, the TL1 function behaves more like the rank function. In this sense, TL1 serves as a valuable theoretical tool to enhance our understanding of the performance between rank minimization and nuclear norm minimization.

Moreover, TL1 is a practical tool, offering significant improvements in accuracy over other methods, regardless of whether it is under uniform or non-uniform sampling with noisy data. The effectiveness of TL1 regularization is demonstrated through a comprehensive simulation study under various missing data mechanisms, highlighting TL1 regularization's adaptability and robustness. We also validate its performance using two real-world data applications, providing empirical evidence of its superiority in practical scenarios. Furthermore, the role of the hyper-parameter a is examined empirically through numerical experiments, which reveals the trade-off between error and rank estimation.

To the best of our knowledge, this work is the first to provide a statistically theoretical analysis of error bounds for TL1 regularization in matrix completion. In fact, it is the pioneering theoretical analysis of a nonconvex method for matrix completion, shedding light on the analysis of other nonconvex regularizations for matrix completion. Additionally, we want to emphasize that it is technically challenging to establish our theorems, as some techniques and properties associated with convex approaches, such as the nuclear norm (Candes and Plan, 2010; Klopp, 2014) and max-norm (Cai and Zhou, 2016), are not directly applicable to the nonconvex TL1 method. And yet, our results are on par with those in convex scenarios. Specifically, Theorems 1 and 2 are minimax optimal (up to a logarithmic factor) in terms of the order with the current literature. Empirically, we demonstrate through extensive experiments that TL1 generally achieves significant improvements over its convex counterparts (including the nuclear norm, max-norm, and a hybrid approach) across various noise levels and sampling distributions.

2 PRELIMINARIES

2.1 Problem Setup

We aim to reconstruct an underlying matrix $A_0 \in \mathbb{R}^{m_1 \times m_2}$ from its partial entries coded by a set of matrices $T_i \in \mathbb{R}^{m_1 \times m_2}$, $i = 1, \ldots, n$, which are i.i.d. copies of a random indicator matrix with distribution $\Pi = (\pi_{kl})_{k,l=1}^{m_1,m_2}$ over the set:

$$\Gamma = \{e_k(m_1)e_l^\top(m_2), k \in [m_1], l \in [m_2]\}, \qquad (1)$$

where π_{kl} is the probability that a particular sample is at the location (k,l), $e_k(m_j)$ represents the canonical basis vector in $\mathbb{R}^{m_j}$ whose k-th entry is 1 and other entries are zeroes, $[m_j] = \{1, \ldots, m_j\}$ for $j = 1, 2$, and n denotes the number of observed samples. In other word, we observe n independent pairs (T_i, Y_i) with $T_i \in \Gamma$ and $Y_i \in \mathbb{R}$ that adhere to the trace regression model (Negahban and Wainwright, 2011)

$$Y_i = \mathrm{tr}(T_i^\intercal A_0) + \sigma\xi_i = \langle T_i, A_0\rangle + \sigma\xi_i, \qquad (2)$$

where $\sigma > 0$ denotes the standard deviation and ξ_i are independent noise variables with $\mathbb{E}(\xi_i) = 0$ and $\mathbb{E}(\xi_i^2) = 1$. Our goal is to estimate the matrix A_0 from observations Y_i and T_i, $i = 1, \ldots, n$.

2.2 Notations

We introduce the notations to be used throughout this paper. For a matrix $A \in \mathbb{R}^{m_1 \times m_2}$, we define constants $m = \min\{m_1, m_2\} = (m_1 \wedge m_2)$, $M = \max\{m_1, m_2\} = (m_1 \vee m_2)$, and $d = m_1 + m_2$. We denote the trace of the matrix A by $\mathrm{tr}(A)$. We define two standard matrix norms, i.e.,

$$\|A\|_\infty = \max_{k,l}|A(k,l)| \text{ and } \|A\|_F = \sqrt{\sum_{k,l} A^2(k,l)},$$

where $A(k,l)$ denotes the value of (k,l)-th entry of A. Denote $\sigma_j(A)$ as the jth singular values of A in decreasing order, then the nuclear norm is defined as $\|A\|_* = \sum_{j=1}^m \sigma_j(A)$ and the spectral norm $\|A\| = \sigma_1(A)$. Given the sampling distribution Π, we define $L_2(\Pi)$ norm of A by $\|A\|_{L_2(\Pi)}^2 = \mathbb{E}(\langle A, T\rangle^2) = \sum_{k=1}^{m_1}\sum_{l=1}^{m_2} \pi_{kl} A^2(k,l)$.

Lastly, our analysis requires the following asymptotic notations. For two non-negative sequences $\{a_n\}$ and $\{b_n\}$, we say $a_n = \mathcal{O}(b_n)$ if there exists a constant C such that $a_n \le Cb_n$ and $a_n = \mathcal{O}_p(b_n)$ if there exists a constant C' such that $a_n \le C'b_n$ with high probability; $a_n = o(b_n)$ if there is a constant C'' such that $a_n < C''b_n$. We denote $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$.

2.3 Model Formulation

TL1 regularization has been used to promote the sparsity in signal recovery (Zhang and Xin, 2018), which can be applied to the singular values of a matrix for low rankness (Zhang et al., 2017). Specifically, the TL1 on the matrix A is defined by

$$\mathrm{TL1}_a(A) = \sum_{j=1}^m \frac{(a+1)\sigma_j(A)}{a + \sigma_j(A)}, \qquad (3)$$

with an internal parameter $a \in (0, \infty)$. The behavior of the TL1 function varies significantly with the parameter a. Specifically, the function has two useful limits:

$$\lim_{a\to 0+} \mathrm{TL1}_a(A) = \mathrm{rank}(A), \quad \lim_{a\to\infty} \mathrm{TL1}_a(A) = \|A\|_*, \qquad (4)$$

allowing the TL1 penalty to act as a bridge between the L_0 semi-norm, which counts non-zero singular values to measure matrix rank directly, and the nuclear norm, which sums singular values as a convex relaxation of the rank. Unlike the L_0 semi-norm, the TL1 penalty with

$a > 0$ is continuous everywhere, which is beneficial for optimization. Since TL1 imposes a heavier penalty on smaller singular values, pushing them toward zero more aggressively than the nuclear norm does, it can induce lower-rank solutions than the nuclear norm, particularly at smaller values of a. However, its non-convex nature poses challenges because non-convex functions typically have multiple local minima and possibly saddle points, complicating the search for a global minimum.

To estimate the target matrix A_0, we incorporate the TL1 function into a least-squares fit of the trace regression, thus leading to

$$\hat{A} = \operatorname*{arg\,min}_{\|A\|_\infty \le \zeta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle T_i, A \rangle)^2 + \lambda \mathrm{TL1}_a(A) \right\}, \tag{5}$$

where $\lambda > 0$ balances between two terms and ζ is a (tunable) upper bound on the estimator.

This paper focuses on the theoretical analysis of the estimator $\hat{A}$ defined in (5). In particular, we aim to find a non-asymptotic upper bound to quantify the difference between the estimated matrix $\hat{A}$ and the true matrix A_0, measured by the Frobenius norm. Our analysis requires $\|\hat{A}\|_\infty \le \zeta$, while ζ can be determined in practice, e.g., $\zeta = 5$ for Netflix ratings in the range of $[0, 5]$. To make the paper self-contained, we include an efficient algorithm for minimizing (5) via the alternating direction method of multipliers (ADMM) (Boyd et al., 2011) in Appendix A.

3 THEORETICAL PROPERTIES

Define $C_l = \sum_{k=1}^{m_1} \pi_{kl}$ as the probability that an observation appears in the l-th column and $R_k = \sum_{l=1}^{m_2} \pi_{kl}$ as the probability that an observation appears in the k-th row, where $k \in [m_1]$ and $l \in [m_2]$. By the definitions along with the constraints $\sum_{l=1}^{m_2} C_l = 1$ and $\sum_{k=1}^{m_1} R_k = 1$, we have $\max_l C_l \ge 1/m_2$ and $\max_k R_k \ge 1/m_1$, implying that $\max_{k,l}(R_k, C_l) \ge 1/m$.

Our theoretical analysis requires the following three assumptions:

Assumption 1. *There exists a constant $L \ge 1$ such that the maximum of R_k and C_l over k and l has an upper bound: $\max_{k,l}(R_k, C_l) \le L/m$.*

Assumption 2. *There exists a constant $\nu \ge 1$ such that $1/(\nu m_1 m_2) \le \pi_{kl} \le \nu/(m_1 m_2)$.*

Assumption 3. *There exists a constant $c_0 > 0$ such that $\max_{i=1,\dots,n} \mathbb{E}[\exp(|\xi_i|/c_0)] \le e$, where e is the base of the natural logarithm.*

These assumptions are commonly used in the literature (Koltchinskii et al., 2011; Klopp, 2014; Cai and Zhou, 2016; Klopp et al., 2017). In Assumption 1, a larger value of L indicates greater imbalance in sampling, leading to a non-uniform setting. Assumption 2 implies that every matrix entry has a non-zero probability of being observed and we have $1/(\nu m_1 m_2)\|A\|_F^2 \le \|A\|_{L_2(\Pi)}^2 \le \nu/(m_1 m_2)\|A\|_F^2$, where ν is a constant independent of n or the matrix dimensions.

Therefore, these two assumptions maintain a balance in the influence exerted by each row and column, ensuring that every matrix element has a chance of being observed, thereby avoiding issues where certain data points are never seen. For a uniform distribution of observed elements, L and ν are taken as 1. Assumption 3 specifies the sub-exponential distribution of the noise term, which is a mild assumption on the noise.

As revealed in Section 2.3, TL1 regularization interpolates between the rank and the nuclear norm of matrices depending on the value of a. In the following, we discuss the theoretical properties of the estimator within two regimes: Regime 1 is when a is large, while Regime 2 is when a is small.

3.1 Regime 1: when a is large

We demonstrate in Theorems 1 and 2 that the estimator obtained by TL1 regularization in (5) achieves the same convergence rate as that obtained by nuclear norm regularization, which implies that the TL1 regularization behaves like the nuclear norm when a is asymptotically large, i.e., $a^{-1} = \mathcal{O}((\zeta\sqrt{m_1 m_2})^{-1})$.

We begin with Theorem 1 to establish an upper bound for the estimation error when the true matrix A_0 is approximately low-rank in the sense that the nuclear norm of A_0 is properly controlled.

Theorem 1. *Suppose Assumptions 1-3 hold, $A_0 \in \mathbb{R}^{m_1 \times m_2}$ is approximately low-rank in the sense that $\|A_0\|_*/\sqrt{m_1 m_2} \le \gamma$ for a constant $\gamma > 0$, and $\|A_0\|_\infty \le \zeta$ for a constant ζ. Take $\lambda \asymp \frac{(\zeta \vee \sigma)}{\sqrt{m_1 m_2}} \frac{a + \zeta\sqrt{m_1 m_2}}{1+a} \sqrt{\frac{Ld \log d}{n}}$, where $d = m_1 + m_2$, then for any $n \gtrsim d \log d$ and $a^{-1} = \mathcal{O}((\zeta\sqrt{m_1 m_2})^{-1})$, there exist two constants C_1 and C_2 only depending on c_0 such that the estimator $\hat{A}$ from (5) satisfies*

$$\frac{1}{m_1 m_2}\|\hat{A} - A_0\|_F^2 \le C_1 \nu(\zeta \vee \sigma)\gamma \sqrt{\frac{Ld \log d}{n}} + C_2 \nu \zeta^2 \sqrt{\frac{L \log d}{n}}, \tag{6}$$

with probability at least $1 - (\kappa + 1)/d$, where κ is a constant depending on L.

Please refer to Appendix B for the proof of Theorem 1.

When $n \gtrsim d \log d$, the first component in (6) dominates in our bound and it is comparable to the result for nearly low-rank matrices in Negahban and Wainwright (2012). Though the second component with order $\mathcal{O}(\sqrt{(\log d)/n})$ is slightly discrepant with a higher order term $(\mathcal{O}(n^{-1}))$ compared to Negahban and Wainwright (2012), it is negligible. Additionally, our bound reaches the minimax optimal rate up to a logarithmic order under the uniform sampling (Cai and Zhou, 2016).

When A_0 is exactly low-rank, Theorem 2 shows that our estimator can achieve a tighter bound than the one in Theorem 1. This is the same situation as the estimation error for nuclear norm regularization (Koltchinskii et al., 2011; Klopp, 2014). However, the proof for the case of the nuclear norm (Negahban and Wainwright, 2012; Klopp, 2014) is not applicable for TL1 regularization, because, unlike the convex nuclear norm, the triangle inequality does not hold for TL1. Instead, we carefully analyze the gradient of the TL1 function to obtain the bound; please refer to Appendix C for the proof of Theorem 2.

Theorem 2. *Suppose Assumptions 1-3 hold, $A_0 \in \mathbb{R}^{m_1\times m_2}$ is exactly low-rank, i.e., $\operatorname{rank}(A_0) \le \tau$ for an integer τ, and $\|A_0\|_\infty \le \zeta$ for a constant ζ. Take $\lambda \asymp \frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}$, where $d = m_1 + m_2$, then for any $n \gtrsim d\log d$ and $a^{-1} = \mathcal{O}((\zeta\sqrt{m_1m_2})^{-1})$, there exist two constants C_3 and C_4 only depending on c_0 such that the estimator $\hat{A}$ from (5) satisfies*

$$\frac{1}{m_1m_2}\|\hat{A}-A_0\|_F^2 \le C_3\nu^2(\zeta^2\vee\sigma^2)\operatorname{rank}(A_0)\frac{Ld\log d}{n} + C_4\nu\zeta^2\sqrt{\frac{L\log d}{n}}, \quad (7)$$

with probability at least $1-(\kappa+1)/d$, where κ is a constant depending on L.

The first component in (7) performs on par with the work (Negahban and Wainwright, 2012). Furthermore, it dominates the second component if $n \le d^2$, which is a mild condition given that $n \ll m_1m_2$ and $d = m_1+m_2$. The upper bound we derive in (7) for exactly low-rank matrices achieves the optimal convergence rate in a minimax sense up to a logarithmic factor in line with the existing literature (Candes and Plan, 2010; Negahban and Wainwright, 2012; Klopp, 2014).

3.2 Regime 2: when a is small

When a approaches 0, the TL1 function approximates the rank function, as indicated in (4). This behavior enhances the ability of TL1 regularization to control the rank of the estimated matrix, which is particularly useful in applications involving low-rank structures. We construct a theoretical foundation to estimate an upper bound of errors in recovering exactly low-rank matrices when a is sufficiently small. We start with a non-asymptotic error bound of exactly low-rank matrices for any value of a that falls outside the range specified in Regime 1, i.e., $a = \mathcal{O}(\zeta\sqrt{m_1m_2})$ for Regime 2.

Theorem 3. *Suppose Assumptions 1-3 hold, $A_0 \in \mathbb{R}^{m_1\times m_2}$ is exactly low-rank, i.e., $\operatorname{rank}(A_0) \le \tau$ for an integer τ, and $\|A_0\|_\infty \le \zeta$ for a constant ζ. Take $\lambda^{-1} = \mathcal{O}\left(\left(\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}\right)^{-1}\right)$, where $d = m_1 + m_2$, then for any $n \gtrsim d\log d$ and $a = \mathcal{O}(\zeta\sqrt{m_1m_2})$, there exist two constants C_5 and C_6 only depending on c_0 such that the estimator $\hat{A}$ from (5) satisfies*

$$\frac{1}{m_1m_2}\|\hat{A}-A_0\|_F^2 \le C_5\lambda\nu\frac{(1+a)\zeta\sqrt{m_1m_2}}{a+\zeta\sqrt{m_1m_2}}\operatorname{rank}(A_0) + C_6\nu\zeta^2\sqrt{\frac{L\log d}{n}}, \quad (8)$$

with probability at least $1-(\kappa+1)/d$, where κ is a constant depending on L.

Please refer to Appendix B for the proof.

Note that even if λ is chosen to be of the exact order in that condition, the error bound in (8) is not as tight as the one of Theorem 2. A specific error bound is presented in Corollary 1 with appropriate choices of λ and a. It is unclear whether this upper bound can be further sharpened, which will be left as a direction for future research. However, by selecting a smaller a, we can gain some control over the rank of the estimator in Theorem 4.

Theorem 4 (Low-rankness)**.** *Suppose Assumptions 1-3 hold, $A_0 \in \mathbb{R}^{m_1\times m_2}$ is exactly low-rank, i.e., $\operatorname{rank}(A_0) \le \tau$ for an integer τ, and $\|A_0\|_\infty \le \zeta$ for a constant ζ. Take $\lambda^{-1} = \mathcal{O}\left(\left(\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}\right)^{-1}\right)$, where $d = m_1 + m_2$, then for any $n \gtrsim d\log d$ and $a = \mathcal{o}((m_1m_2)^{1/4})$, there exists a constant C_7 only depending on c_0 such that the estimator $\hat{A}$ from (5) satisfies*

$$\operatorname{rank}(\hat{A}) \le C_7\Bigg\{\lambda^{-1}\frac{Ld\log d\sqrt{m_1m_2}}{(1+a)(a+\sqrt{m_1m_2})n} \times\left(\sqrt{a}/(a+\sqrt{m_1m_2})^{1/4}+1\right)^2 + \operatorname{rank}(A_0)\left(\sqrt{a}/(a+\sqrt{m_1m_2})^{1/4}+1\right)\Bigg\}, \quad (9)$$

with high probability.

Theorem 4 provides a theoretical control on the rank of the matrix estimated by the TL1 regularization, showing the estimated rank decreases as λ increases for sufficiently small a. Theorem 4 indicates that TL1 regularization controls the rank of estimated matrices to some extent and implies that TL1 regularization effectively promotes sparsity in the singular values. To the best of our knowledge, there is no comprehensive analysis in the current literature that bounds the rank of estimated matrices by nuclear norm or max-norm.

Furthermore, there is a trade-off between controlling the estimation error and the rank. For a fixed value of a, a larger value of λ makes the first component in the bound (9) smaller, thus resulting in a lower rank estimation. However, a lower rank estimator comes at the cost of a slower convergence rate of the errors in the Frobenius norm, compared to the rate specified in Theorem 3. This trade-off underlines the balance between achieving a low-rank representation and maintaining a high convergence rate (or minimizing error estimation). In addition, smaller a makes the second component in (9) smaller and yields a lower rank estimation, highlighting the importance of carefully choosing the hyper-parameters λ and a to optimize both rank reduction and error minimization in a balanced manner.

Combining Theorems 3 and 4, we present a specific upper bound for the estimation error and the rank with appropriate choices of λ and a in Corollary 1. The proofs of Theorem 4 and Corollary 1 can be found in Appendix D.

Corollary 1. *Suppose Assumptions 1-3 hold, $A_0 \in \mathbb{R}^{m_1 \times m_2}$ is exactly low-rank, i.e.,* $\text{rank}(A_0) \leq \tau$ *for an integer τ. Take $\lambda \asymp \frac{(\zeta \vee \sigma)}{\sqrt{m_1 m_2}} \frac{a+\zeta\sqrt{m_1 m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}$, where $d = m_1 + m_2$, then for any $n \gtrsim d \log d$ and $a = \mathcal{o}((m_1 m_2)^{1/4})$, there exist two constants C_8 and C_9 only depending on c_0 such that the estimator $\hat{A}$ from* (5) *satisfies*

$$\frac{1}{m_1 m_2}\|\hat{A}-A_0\|_F^2 \leq C_8 \nu(\zeta^2 \vee \sigma^2)\text{rank}(A_0)\sqrt{\frac{Ld\log d}{n}} + C_9 \nu \zeta^2 \sqrt{\frac{L\log d}{n}}, \quad (10)$$

with probability at least $1-(\kappa+1)/d$. Using the condition that $d \log d = \mathcal{O}(n)$, we have

$$\text{rank}(\hat{A}) = \mathcal{O}_p(\text{rank}(A_0)). \quad (11)$$

Corollary 1 provides a concrete upper bound for error estimation in the Frobenius norm and establishes a bound on the rank of the estimator by selecting appropriate values of λ and a. When the sample size n increases, the upper bound of the error decreases at a rate of $(d \log d/n)^{1/2}$, and the rank remains in a constant order, independent of the size of the matrix. This constant order bound of the rank is significantly better than the worst-case bound, which is $\mathcal{O}(m)$. This result is aligned with the rank of the true matrix that is bounded by a constant. Additionally, it is worth mentioning that our proof techniques, detailed in Appendix D, can be utilized for other non-convex regularizations, such as SCAD regularization (Fan and Li, 2001), MCP regularization (Zhang, 2010), L_p norm (Chartrand, 2007), and L_1/L_2 functional (Rahimi et al., 2019) on the singular values of matrices. One limitation of Corollary 1 is that (11) does not imply the ideal scenario when $\text{rank}(\hat{A}) = \text{rank}(A_0)$. In future work, we will investigate whether this oracle property can be achieved by properly choosing a and λ.

4 SIMULATION STUDY

In this simulation study, we generate the target matrix $A_0 \in \mathbb{R}^{m_1 \times m_2}$ as the product of two matrices of smaller dimensions, i.e., $A_0 = UV^\intercal$, where $U \in \mathbb{R}^{m_1 \times r}$, $V \in \mathbb{R}^{m_2 \times r}$ with each entry of U and V independently sampled from a standard normal distribution $\mathcal{N}(0, 1)$. As a result, the rank of A_0 is at most r, which is significantly smaller than $\min(m_1, m_2)$.

We adopt three sampling schemes outlined in Fang et al. (2018) to facilitate a direct comparison of completion results. All the sampling schemes are implemented without replacement. The first scheme involves uniform sampling of the indices of observed entries, while the subsequent two schemes sample indices in a non-uniform manner. Specifically, for each $(k, l) \in [m_1] \times [m_2]$, let p_k (and p_l) be:

Scheme 1	Scheme 2	Scheme 3
$p_k =$ $\frac{1}{m_1 m_2}$	$p_k =$ $\begin{cases} 2p_0 \text{ if } k \leq \frac{m_1}{10} \\ 4p_0 \text{ if } \frac{m_1}{10} < k \leq \frac{m_1}{5} \\ p_0 \quad \text{otherwise} \end{cases}$	$p_k =$ $\begin{cases} 3p_0 \text{ if } k \leq \frac{m_1}{10} \\ 9p_0 \text{ if } \frac{m_1}{10} < k \leq \frac{m_1}{5} \\ p_0 \quad \text{otherwise} \end{cases}$

where p_0 is a normalized constant such that $\sum_{k=1}^{m_1} p_k = 1$. Let $\pi_{kl}{=}p_k p_l$, then the sampling distribution is $\Pi = (\pi_{kl})_{k,l=1}^{m_1,m_2}$. Note that the configuration of Scheme 3 demonstrates a higher level of imbalance, as some entries have higher probabilities, meaning they are more likely to be observed than others. This imbalance poses a greater challenge for recovery efforts compared to Scheme 1 and Scheme 2, as evidenced by more substantial errors in Table 3 than Tables 1 and 2. Please refer to our Github: `https://github.com/Kun9550/NoisyLRMC_via_TL1`, regarding a demo code for these sampling schemes along with our TL1 implementation.

From the trace regression model (2), we define sampling ratios as $\text{SR} = n/(m_1 m_2)$ and set the value of σ such that the signal-to-noise ratio, defined by $\text{SNR} = 10\text{log}10(\frac{1}{\sigma^2}\sum_i^n \langle T_i, A_0\rangle^2)$, is either 10 or 20. To quantitatively evaluate the matrix recovery performance, we use the relative error (RE) defined by $\text{RE} = \|\hat{A} - A_0\|_F / \|A_0\|_F$, where $\hat{A}$ is an estimator of A_0. To find $\hat{A}$, we compare the TL1-regularized model (5) with nuclear norm (Candes and Recht, 2012), max-norm (Cai and Zhou, 2016), and a hybrid approach combining both the nuclear norm and max-norm (Fang et al., 2018). We explore various combinations of dimensions, ranks, sampling schemes, and sampling ratios (SR) under noisy cases. For each combination, we select the optimal parameters for each competing method; please refer to Appendix E for more details on parameter tuning. Due to space limitations, we only display the results for SNR = 10 here. Please refer to Appendix F for the results of SNR = 20.

We present the recovery comparison under the three sampling schemes in Tables 1, 2, and 3, respectively. In Table 1 where uniformly sampled data is used, our results show that the TL1 regularization method yields the most favorable recovery outcomes in a noisy setting. Notably, the hybrid approach outperforms the max-norm method, aligning with established numerical analyses in Fang et al. (2018) that posit the inferior performance of the max-norm method compared to the nuclear norm when the observed entries are indeed uniformly sampled. In Tables 2 and 3 where non-uniform sampling distributions are employed, we find that TL1 regularization performs best with a few exceptions. Furthermore, the results in Table 3 are generally worse than the ones in Table 2, which indicates the recovery difficulty of Scheme 3 for matrix completion. In short, we conclude that the TL1-regularized approach is robust across various noise levels and sampling distributions.

Discussion 1. We explore how the hyper-parameter a in the TL1 regularization affects the performance of the model (5) for low-rank matrix completion. In particular, we examine the rank estimations of different methods for a 500×500 target matrix of rank 5 under Scheme 2 with 20% sampling rate and SNR = 10 in one simulation. We set the hyper-parameter $a = 100$ for TL1 regularization, which is mostly chosen among all the simulation scenarios. It is evident in Table 4 that the TL1-regularized model results in the lowest rank for the estimator while keeping the smallest relative error.

Figure 1: Impact of the parameter a in TL1 on the matrix recovery: relative errors (left) and estimated rank (right) with respect to a.

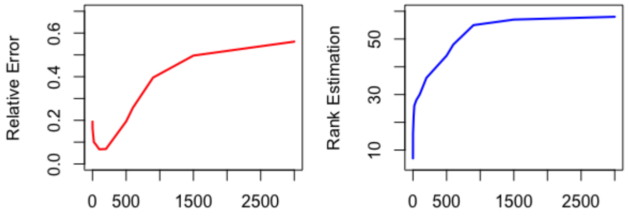

Figure 1 illustrates the changes in the relative error and the rank of the estimator matrix with respect to the parameter a. For each a, we select the optimal value of λ in the model (5). We observe in Figure 1 that the relative error decreases as a increases from 0, reaching a minimum around $a = 100$, and then increases steadily, eventually becoming asymptotically flat. For the rank estimation, there is an increasing trend as the value of a increases. The curve rises sharply at first, indicating that smaller values of a lead to better rank estimation. When a is relatively large, the estimated rank becomes close to that of the nuclear norm method. This empirical behavior aligns with the TL1's property in (4). In Table 4, we present the rank estimated by all competing methods under the same setting, showing that the TL1 regularization produces a matrix with the lowest rank among them. Although the rank estimated by TL1 regularization is higher than the true rank (i.e., 5), this outcome is reasonable due to the trade-off between estimation error and rank.

Discussion 2. We compare the bias and variance of estimators obtained from TL1 regularization and nuclear norm regularization. The TL1-regularized model is expected to introduce less bias than nuclear norm regularization, as TL1 reduces the penalization on larger singular values, potentially leading to higher accuracy in matrix completion. In contrast, the nuclear norm-based model minimizes the sum of singular values, which generally reduces the estimator's variance but tends to introduce higher bias. To empirically validate these hypotheses, we design an experiment under Scheme 2 with SNR= 10 over 100 random trials. The results, presented in Table 9 in Appendix F, show that the estimator from the TL1-regularized model exhibits a lower bias but slightly higher variance compared to the nuclear norm regularized model.

5 REAL DATA APPLICATIONS

We investigate the performance of two real-world datasets: Coat Shopping and MovieLens 100K. Please refer to Appendix G for data description. For both datasets, we randomly divide the test set into two distinct subsets: a validation set for the purpose of tuning parameters and an evaluation set to assess the performance of the estimator. Each subset includes 50 percent of the observed entries from the original test set. We use the test root mean squared error (TRMSE) restricted on the evaluation set to gauge the recovery performance, as outlined in Wang et al. (2021). We report the TRMSE values and the ranks of the estimators in Table 5, showing that TL1 outperforms the other approaches in terms of TRMSE for both datasets. For the Coat Shopping Dataset, the hybrid-regularized method provides the second-smallest TRMSE, but it produces a higher rank than the TL1-based method. In contrast, the nuclear norm regularized model performs the worst.

Table 1: Relative errors of the reconstructed matrix to the ground truth under Scheme 1 setting with SNR=10. Each reported value is the average RE of over 50 random realizations, with standard deviation in parentheses. We highlight the best values (smallest REs) using boldface and 2nd best in italics.

	(r, SR)	Max-norm		Hybrid		Nuclear		TL1	
		RE	Time	RE	Time	RE	Time	RE	Time
300	(5, 0.1)	0.350 (0.011)	16.14	0.217 (0.008)	17.24	*0.214 (0.008)*	7.82	**0.076 (0.002)**	7.15
	(5, 0.2)	0.200 (0.005)	19.64	0.095 (0.002)	22.10	*0.085 (0.001)*	6.32	**0.045 (0.001)**	7.45
	(10, 0.1)	0.585 (0.011)	24.12	*0.531 (0.011)*	24.07	0.534 (0.010)	6.54	**0.161 (0.003)**	7.84
	(10, 0.2)	0.260 (0.006)	20.97	*0.157 (0.003)*	23.06	0.159 (0.003)	6.56	**0.071 (0.001)**	7.28
500	(5, 0.1)	0.225 (0.007)	49.79	0.118 (0.002)	50.02	*0.115 (0.002)*	25.82	**0.076 (0.001)**	28.77
	(5, 0.2)	0.134 (0.004)	48.18	0.079 (0.000)	58.31	*0.073 (0.001)*	25.88	**0.068 (0.001)**	29.18
	(10, 0.1)	0.324 (0.015)	49.21	0.244 (0.005)	49.96	*0.242 (0.005)*	25.77	**0.092 (0.001)**	29.18
	(10, 0.2)	0.164 (0.003)	48.32	0.103 (0.001)	59.25	*0.098 (0.001)*	25.95	**0.081 (0.001)**	29.33

Table 2: Relative errors of the reconstructed matrix to the ground truth under Scheme 2 with SNR=10.

	(r, SR)	Max-norm		Hybrid		Nuclear		TL1	
		RE	Time	RE	Time	RE	Time	RE	Time
300	(5, 0.1)	0.347 (0.004)	22.57	**0.338 (0.003)**	21.02	0.767 (0.002)	4.79	*0.402 (0.006)*	8.40
	(5, 0.2)	0.213 (0.002)	24.68	*0.193 (0.003)*	24.56	0.611 (0.003)	5.66	**0.114 (0.003)**	8.50
	(10, 0.1)	0.521 (0.003)	25.72	**0.517 (0.003)**	21.94	0.806 (0.001)	4.96	*0.560 (0.003)*	7.87
	(10, 0.2)	0.265 (0.002)	24.01	*0.264 (0.003)*	17.06	0.620 (0.002)	6.79	**0.188 (0.003)**	8.15
500	(5, 0.1)	0.282 (0.004)	59.46	*0.251 (0.003)*	51.13	0.758 (0.002)	19.23	**0.106 (0.002)**	22.11
	(5, 0.2)	0.161 (0.001)	59.66	*0.149 (0.001)*	55.75	0.604 (0.002)	21.55	**0.066 (0.000)**	24.36
	(10, 0.1)	0.365 (0.002)	57.83	*0.331 (0.002)*	50.83	0.771 (0.001)	19.56	**0.197 (0.003)**	22.31
	(10, 0.2)	0.201 (0.000)	58.72	*0.181 (0.000)*	57.62	0.613 (0.001)	21.51	**0.087 (0.001)**	24.02

Table 3: Relative errors of the reconstructed matrix to the ground truth under Scheme 3 with SNR=10.

	(r, SR)	Max-norm		Hybrid		Nuclear		TL1	
		RE	Time	RE	Time	RE	Time	RE	Time
300	(5, 0.1)	0.488 (0.018)	18.92	**0.487 (0.020)**	17.29	0.790 (0.010)	5.78	*0.490 (0.027)*	6.35
	(5, 0.2)	0.219 (0.018)	19.53	*0.207 (0.019)*	17.98	0.617 (0.016)	6.60	**0.122 (0.021)**	7.11
	(10, 0.1)	0.650 (0.015)	19.19	*0.655 (0.016)*	16.97	0.833 (0.007)	5.80	**0.572 (0.016)**	6.40
	(10, 0.2)	0.279 (0.016)	19.28	*0.267 (0.018)*	17.72	0.638 (0.011)	6.58	**0.208 (0.024)**	7.12
500	(5, 0.1)	0.406 (0.023)	63.43	*0.350 (0.023)*	61.82	0.764 (0.013)	60.88	**0.171 (0.026)**	60.26
	(5, 0.2)	0.177 (0.008)	53.47	*0.159 (0.007)*	50.34	0.611 (0.014)	50.16	**0.098 (0.001)**	50.45
	(10, 0.1)	0.522 (0.011)	19.34	*0.493 (0.012)*	20.54	0.798 (0.007)	19.12	**0.380 (0.015)**	20.81
	(10, 0.2)	0.296 (0.007)	22.53	*0.214 (0.009)*	23.68	0.657 (0.010)	22.05	**0.128 (0.005)**	23.53

Table 4: The rank estimation while keeping the smallest error for different approaches under Scheme 2 with SNR=10.

$m_1 = m_2$	(r, SR)	Max-norm	Hybrid	Nuclear	TL1
500	(5, 0.2)	42	36	58	28

Table 5: Comparison in terms of TRMSE and estimated rank on the Coat Shopping Dataset and MovieLens 100K Dataset. We highlight the best TRMSE values using boldface and 2nd best in italics.

	Coat Shopping Dataset				MovieLens 100K Dataset			
Method	Max-norm	Hybrid	Nuclear	TL1	Max-norm	Hybrid	Nuclear	TL1
TRMSE	1.0635	*1.0107*	1.1726	**0.9975**	*1.0128*	1.0375	1.1586	**1.0051**
Rank	28	30	33	22	38	35	59	30

For the MovieLens dataset, TL1 regularization delivers the best TRMSE with the lowest rank. In summary, the TL1-regularized model demonstrates significant improvement over other methods with consistent smallest errors and relatively lower rank estimation.

6 CONCLUSION

This paper makes a contribution to the statistical analysis for controlling the rank of the estimator and the non-asymptotic upper bounds of recovery errors produced by the TL1 regularization. For an asymptotically large parameter a, the error bound achieves a minimax optimal convergence rate up to a logarithmic factor for low-rank or approximately low-rank matrices, aligned with existing literature work on nuclear norm regularization. Regarding the theoretical analysis of rank estimation which has not yet been extensively studied, we establish a constant order bound for exactly low-rank matrices when a is small, and we aim to refine this bound in future research. Overall, our work bridges the gap in the literature, which primarily addresses matrix completion with uniform sampling, by extending it to a more general sampling mechanism. Experimental results demonstrate that TL1 outperforms other methods with consistently smaller errors and lower ranks regardless of whether the sampling scheme is uniform. We include an empirical study of the parameter a for its impact on the matrix completion, aiming at some guidance in the practical use of TL1 regularize.

Acknowledgments

The authors thank the reviewers for their helpful comments and suggestions. Portions of this research were conducted with the computing resources provided by UTD High Performance Computing (HPC). The work of Jiayi Wang is partly supported by the National Science Foundation (DMS-2401272). Yifei Lou is partly supported by the NSF CAREER Award 2414705.

References

Kwangjun Ahn and Felipe Suarez. Riemannian perspective on matrix factorization. *arXiv preprint arXiv:2102.00937*, 2021.

James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, 2007.

Sonia A Bhaskar. Probabilistic low-rank matrix completion from quantized measurements. *Journal of Machine Learning Research*, 17(60):1–34, 2016.

Xuan Bi, Annie Qu, Junhui Wang, and Xiaotong Shen. A group-specific recommender system. *Journal of the American Statistical Association*, 112(519):1344–1353, 2017.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4): 1956–1982, 2010.

T Tony Cai and Wen-Xin Zhou. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1):1493–1525, 2016.

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

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

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

Feilong Cao, Miaomiao Cai, and Yuanpeng Tan. Image interpolation via low-rank matrix completion and recovery. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(8):1261–1270, 2014.

Fu Changjun, Ji Xiangyang, Zhang Yongbing, and Dai Qionghai. A single frame super-resolution method based on matrix completion. In *2012 Data Compression Conference*, pages 297–306. IEEE, 2012.

Rick Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.

Ji Chen, Xiaodong Li, and Zongming Ma. Nonconvex matrix completion with linearly parameterized factors. *Journal of Machine Learning Research*, 23 (207):1–35, 2022.

Junren Chen, Cheng-Long Wang, Michael K Ng, and Di Wang. High dimensional statistical estimation under uniformly dithered one-bit quantization. *IEEE Transactions on Information Theory*, 2023.

Xiaobo Chen, Zhongjie Wei, Zuoyong Li, Jun Liang, Yingfeng Cai, and Bob Zhang. Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation. *Knowledge-Based Systems*, 132:249–262, 2017.

Xinyu Chen, Jinming Yang, and Lijun Sun. A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation. *Transportation Research Part C: Emerging Technologies*, 117:102673, 2020a.

Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Completing any low-rank matrix, provably. *The Journal of Machine Learning Research*, 16(1):2999–3034, 2015.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020b.

Zhaoliang Chen, Wei Zhao, and Shiping Wang. Kernel meets recommender systems: A multi-kernel interpolation for matrix completion. *Expert Systems with Applications*, 168:114436, 2021.

Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain. Nearly optimal robust matrix completion. In *International Conference on Machine Learning*, pages 797–805. PMLR, 2017.

Alexander L Chistov and D Yu Grigor'Ev. Complexity of quantifier elimination in the theory of algebraically closed fields. In *International Symposium on Mathematical Foundations of Computer Science*, pages 17–31. Springer, 1984.

Juhee Cho, Donggyu Kim, and Karl Rohe. Asymptotic theory for estimating the singular vectors and values of a partially-observed low rank matrix with noise. *Statistica Sinica*, pages 1921–1948, 2017.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96 (456):1348–1360, 2001.

Ethan X Fang, Han Liu, Kim-Chuan Toh, and Wen-Xin Zhou. Max-norm optimization for robust matrix recovery. *Mathematical Programming*, 167:5–35, 2018.

Vivek Farias, Andrew A Li, and Tianyi Peng. Uncertainty quantification for low-rank matrix completion with heterogeneous and sub-exponential noise. In *International Conference on Artificial Intelligence and Statistics*, pages 1179–1189. PMLR, 2022.

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

Yuzhou Gu, Zhao Song, Junze Yin, and Lichen Zhang. Low rank matrix completion via robust alternating minimization in nearly linear time. *arXiv preprint arXiv:2302.11068*, 2023.

Davide Feltoni Gurini, Fabio Gasparetti, Alessandro Micarelli, and Giuseppe Sansonetti. Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization. *Future Generation Computer Systems*, 78:430–439, 2018.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on interactive intelligent systems (TIIS)*, 5(4): 1–19, 2015.

Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

Hui Ji, Chaoqiang Liu, Zuowei Shen, and Yuhong Xu. Robust video denoising using low rank matrix completion. In *Conference on Computer Vision and Pattern Recognition*, pages 1791–1798. IEEE, 2010.

Zhigang Jia, Qiyu Jin, Michael K Ng, and Xi-Le Zhao. Non-local robust quaternion matrix completion for large-scale color image and video inpainting. *IEEE Transactions on Image Processing*, 31:3868–3883, 2022.

Huaqing Jin, Yanyuan Ma, and Fei Jiang. Matrix completion with covariate information and informative missingness. *Journal of Machine Learning Research*, 23(180):1–62, 2022.

Zhao Kang, Chong Peng, and Qiang Cheng. Top-n recommender system via matrix completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

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

Franz J Király, Louis Theran, and Ryota Tomioka. The algebraic combinatorial approach for low-rank matrix completion. *Journal of Machine Learning Research*, 16(1):1391–1436, 2015.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 2014.

Olga Klopp, Jean Lafond, Éric Moulines, and Joseph Salmon. Adaptive multinomial matrix completion. *Electronic Journal of Statistics*, 9(2):2950–2975, 2015.

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

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 2011.

Deniz Kumlu. Ground penetrating radar data reconstruction via matrix completion. *International Journal of Remote Sensing*, 42(12):4607–4624, 2021.

Jiangyuan Li, Jiayi Wang, Raymond KW Wong, and Kwun Chuen Gary Chan. A pairwise pseudo-likelihood approach for matrix completion with informative missingness. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Rongrong Ma, Jianyu Miao, Lingfeng Niu, and Peng Zhang. Transformed l1 regularization for learning sparse deep neural networks. *Neural Networks*, 119: 286–298, 2019.

Wei Ma and George H Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Ranald R MacDonald. Missing data–quantitative applications in the social sciences. *British Journal of Mathematical & Statistical Psychology*, 55:193, 2002.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24 (2):227–234, 1995.

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

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13:1665–1697, 2012.

Jonathan Popa, Susan E Minkoff, and Yifei Lou. An improved seismic data completion algorithm using low-rank tensor optimization: Cost reduction and optimal data orientation. *Geophysics*, 86(3):V219–V232, 2021.

Jonathan Popa, Susan E Minkoff, and Yifei Lou. Tensor-based reconstruction applied to regularized time-lapse data. *Geophysical Journal International*, 231(1):638–649, 2022.

Yaghoub Rahimi, Chao Wang, Hongbo Dong, and Yifei Lou. A scale-invariant approach for sparse signal recovery. *SIAM Journal on Scientific Computing*, 41 (6):A3649–A3672, 2019.

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

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning*, pages 1670–1679. PMLR, 2016.

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

Nathan Srebro and Russ R Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 23, 2010.

Nathan Srebro, Jason Rennie, and Tommi Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 17, 2004.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021.

Jiayi Wang, Raymond KW Wong, Xiaojun Mao, and Kwun Chuen Gary Chan. Matrix completion with model-free weighting. In *International Conference on Machine Learning*, pages 10927–10936. PMLR, 2021.

Zhiyuan Weng and Xin Wang. Low-rank matrix completion for array signal processing. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2697–2700. IEEE, 2012.

Dong Xia and Ming Yuan. Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(1):58–77, 2021.

Xiuqin Xu, Mingwei Lin, Xin Luo, and Zeshui Xu. HRST-LR: a hessian regularization spatio-temporal low rank algorithm for traffic data imputation. *IEEE Transactions on Intelligent Transportation Systems*, 24:11001–11017, 2023.

Henry Shaowu Yuchi, Simon Mak, and Yao Xie. Bayesian uncertainty quantification for low-rank matrix completion. *Bayesian Analysis*, 18(2):491–518, 2023.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942, 2010.

Qian Zhang, Jie Lu, and Yaochu Jin. Artificial intelligence in recommender systems. *Complex & Intelligent Systems*, 7(1):439–457, 2021.

Shuai Zhang and Jack Xin. Minimization of transformed l 1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing. *Mathematical Programming*, 169(1):307–336, 2018.

Shuai Zhang, Penghang Yin, and Jack Xin. Transformed schatten-1 iterative thresholding algorithms for low rank matrix completion. *Communications in Mathematical Sciences*, 15:839 – 862, 2017.

Shuimei Zhang and Yimin D Zhang. Low-rank hankel matrix completion for robust time-frequency analysis. *IEEE Transactions on Signal Processing*, 68:6171–6186, 2020.

Huiwen Zheng, Yifei Lou, Guoliang Tian, and Chao Wang. A scale-invariant relaxation in low-rank tensor recovery with an application to tensor completion. *SIAM Journal on Imaging Sciences*, 17(1):756–783, 2024.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   Justification: All the model settings, assumptions and algorithm can be found in the Appendix or main text.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   Justification: We provided the information in the Appendix.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   Justification: All assumptions can be found in the Appendix or main text.
   (b) Complete proofs of all theoretical results. [Yes]
   Justification: All the proofs can be found in the Appendix or main text.
   (c) Clear explanations of any assumptions. [Yes]
   Justification: The clear explanations of assumptions can be found in Section 3.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   Justification: The paper fully discloses all the information needed to reproduce the main experimental results. This includes detailed descriptions of the datasets, scenarios setup, hyper-parameter settings, performance measurements, algorithm, and experimental results.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   Justification: All the details can be found in the Appendix.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   Justification: All the details can be found in the Appendix or main text.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
   Justification: Our experiments are performed on a MacBook Pro with an Apple M2 Chip and 8GB memory.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Nosiy Low-Rank Matrix Completion via Transformed L_1 Regularization and its Theoretical Properties: Supplementary Materials

A Numerical algorithm

Here we describe an efficient algorithm for solving the TL1-regularized model for matrix completion, defined in (5). Recall that we are given n independent pairs (T_i, Y_i) with $T_i \in \Gamma$ and $Y_i \in \mathbb{R}$ in the trace regression model (2), which can be expressed by the Hadamard product, i.e.,

$$Y = T \circ A_0 + N, \tag{12}$$

where $T = \sum_{i=1}^n T_i$, $\circ$ denotes the elementwise Hadamard product, N is the noise term. The resulting matrix Y is of the same size as the underlying matrix $A_0 \in \mathbb{R}^{m_1 \times m_2}$. Consequently, the optimization problem (5) can be reformulated as:

$$\begin{aligned} &\min_{A,Z} \frac{1}{n}\|Y - T \circ A\|_F^2 + \lambda \mathrm{TL1}_a(Z) \\ &\text{subject to } A = Z \text{ and } \|A\|_\infty \leq \zeta, \end{aligned} \tag{13}$$

with an auxiliary variable $Z \in \mathbb{R}^{m_1 \times m_2}$ so that we can apply the alternating direction method of multipliers (ADMM) (Boyd et al., 2011). In particular, the augmented Lagrangian function of (13) can be written as

$$\mathcal{L}(A, Z, W) = \frac{1}{n}\|Y - T \circ A\|_F^2 + \lambda \mathrm{TL1}_a(Z) + \langle A - Z, W \rangle + \frac{\rho}{2}\|A - Z\|_F^2, \tag{14}$$

where $W \in \mathbb{R}^{m_1 \times m_2}$ is a dual variable and $\rho > 0$ is a step size. The ADMM scheme involves iteratively minimizing the augmented Lagrangian with respect to A while keeping Z and W fixed, then minimizing with respect to Z while keeping A and W fixed, and finally performing a gradient ascent step with respect to W while keeping A and Z fixed. In short, ADMM iterates as follows,

$$A^{k+1} = \arg\min_{\|A\|_\infty \leq \zeta} \mathcal{L}(A, Z^k, W^k), \tag{15}$$

$$Z^{k+1} = \arg\min_{Z} \mathcal{L}(A^{k+1}, Z, W^k), \tag{16}$$

$$W^{k+1} = W^k + \tau\rho(A^{k+1} - Z^{k+1}), \tag{17}$$

where $\tau \in \big(0, (1+\sqrt{5})/2\big)$ is a parameter to acceleration. We follow the work of Fang et al. (2018) to set $\tau = 1.618$.

For the A sub-problem (15), we write it as

$$A^{k+1} = \arg\min_{\|A\|_\infty \leq \zeta} \frac{1}{n}\|Y - T \circ A\|_F^2 + \frac{\rho}{2}\|A - Z^k + \frac{1}{\rho}W^k\|_F^2. \tag{18}$$

Without the constraint $\|A\|_\infty \leq \zeta$, the optimal solution to (18) can be expressed by

$$A^{k+\frac{1}{2}} := \left(\frac{2}{n}T \circ Y + \rho Z^k - W^k\right) \oslash \left(\frac{2}{n}T + \rho\right), \tag{19}$$

where $\oslash$ denotes the elementwise division. Then we project the solution to the constraint $[-\zeta, \zeta]$, thus leading to

$$A^{k+1} = \min\left\{\max\{A^{k+\frac{1}{2}}, \zeta\}, -\zeta\right\}, \tag{20}$$

where min and max are conducted elementwise.

The Z subproblem (16) can be formulated by

$$Z^{k+1} = \arg\min_{Z} \lambda \mathrm{TL1}_a(Z) + \frac{\rho}{2}\|X^{k+1} - Z + \frac{1}{\rho}W^k\|_F^2, \tag{21}$$

which has a closed-form solution that is similar to the singular value thresholding (SVT) operator for the nuclear norm minimization (Cai et al., 2010). Specifically, we define the proximal operator (Zhang et al., 2017) for the TL1 regularization applied on a scalar x to be

$$\begin{aligned}\mathrm{prox}_{\mathrm{TL1}_a}(x,\mu) :=& \arg\min_{z\in\mathbb{R}} \left\{ \mu\frac{(a+1)z}{a+z} + \frac{1}{2}(z-x)^2 \right\}\\ =&\mathrm{sign}(x)\left\{ \frac{2}{3}(a+|x|)\cos(\frac{\phi(x)}{3}) - \frac{2a}{3} + \frac{|x|}{3} \right\},\end{aligned}$$

with $\phi(x) = \arccos(1 - \frac{27\mu a(1+a)}{2(a+|x|)^3})$. Then, the closed-form solution for the Z-update (21) is given by

$$Z^{k+1} = U\mathrm{diag}\left(\{\mathrm{prox}_{\mathrm{TL1}_a}(\sigma_k, \lambda/\rho)\}_{1\le k\le m}\right) V^\intercal, \tag{22}$$

where we have the singular value decomposition (SVD) of the matrix $A^{k+1} + \frac{1}{\rho}W^k = U\Sigma V^\intercal$ and the diagonal matrix Σ has elements σ_k for $1 \le k \le m$.

We summarize the overall algorithm in Algorithm 1.

Algorithm 1 TL1-regularized matrix completion (5) via ADMM

```
Input: $Y$, $T$
Set parameters: $\lambda$, $a$, $\zeta$, $\rho$, $\tau = 1.618$
Initialize $Z^0 = Y$, $V^0 = 0$ and $k = 0$.
while stopping criterion is not satisfied do
    $A^{k+1} \leftarrow \min\left\{\max\{\left(\frac{2}{n}T \circ Y + \rho Z^k - W^k\right) \oslash (\frac{2}{n}T + \rho), \zeta\}, -\zeta\right\}$
    $Z^{k+1} \leftarrow U\mathrm{diag}\left(\{\mathrm{prox}_{\mathrm{TL1}_a}(\sigma_k, \lambda/\rho)\}_k\right) V^\intercal$, where $A^{k+1} + \frac{1}{\rho}W^k = U\mathrm{diag}(\{\sigma_k\}_k)V^\intercal$
    $W^{k+1} \leftarrow W^k + \tau\rho(A^{k+1} - Z^{k+1})$
    $k \leftarrow k + 1$
end while
```

B Proof of Theorem 1 and Theorem 3

Denote a constraint set:

$$\mathcal{K}(\zeta,\gamma) := \left\{ A \in \mathbb{R}^{m_1\times m_2} : \|A\|_\infty \le \zeta, \frac{\|A\|_*}{\sqrt{m_1 m_2}} \le \gamma \right\}, \tag{23}$$

and a constant

$$Z_\gamma = \sup_{A\in\mathcal{K}(\zeta,\gamma)} \left| \frac{1}{n}\sum_{i=1}^n \langle T_i, A\rangle^2 - \|A\|^2_{L_2(\Pi)} \right|. \tag{24}$$

Given an i.i.d. Rademacher sequence $\{\epsilon_i\}_{i=1}^n$, we define,

$$\Sigma_R = \frac{1}{n}\sum_{i=1}^n \epsilon_i T_i \text{ and } \Sigma = \frac{\sigma}{n}\sum_{i=1}^n \xi_i T_i. \tag{25}$$

We introduce Lemmas 1-3 to find a non-asymptotic upper bound on the Frobenius norm error.

Lemma 1. *Suppose that T_i are i.i.d. indicator matrices with distribution Π on Γ for $i = 1, \ldots, n$, and Assumptions 1 and 2 hold. For any $n \ge m\frac{(\log d)^3}{L}$, there exists a constant C^* only depending on c_0 such that*

$$\mathbb{E}(Z_\gamma) \le C^*\gamma\sqrt{\frac{LM\log d}{n}},$$

with a probability at least $1 - \frac{1}{d}$.

Proof of Lemma 1. Without loss of generality, we take $\zeta = 1$ in the proposed model (5). Consequently, $\|A\|_\infty \le 1$ indicates $|\langle T_i, A\rangle| \le 1$, $i = 1, \ldots, n$. It follows from the symmetrization inequality from the book (Koltchinskii, 2011, Theorem 2.1) that

$$\mathbb{E}(Z_\gamma) = \mathbb{E}\left(\sup_{A\in\mathcal{K}(1,\gamma)} \left| \frac{1}{n}\sum_{i=1}^n \langle T_i, A\rangle^2 - \|A\|^2_{L_2(\pi)} \right| \right) \le 2\mathbb{E}\left(\sup_{A\in\mathcal{K}(1,\gamma)} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i \langle T_i, A\rangle^2 \right| \right),$$

where $\{\epsilon_i\}_{i=1}^n$ are independent Rademacher random variables. We further use the contraction inequality in Koltchinskii (2011, Theorem 2.3) to obtain that

$$\begin{aligned}\mathbb{E}(Z_\gamma) &\le 8\mathbb{E}\biggl(\sup_{A\in\mathcal{K}(1,\gamma)}\biggl|\frac{1}{n}\sum_{i=1}^n\epsilon_i\langle T_i,A\rangle\biggr|\biggr) = 8\mathbb{E}\biggl(\sup_{A\in\mathcal{K}(1,\gamma)}|\langle\Sigma_R,A\rangle|\biggr)\\ &\le 8\mathbb{E}\biggl(\sup_{A\in\mathcal{K}(1,\gamma)}\|\Sigma_R\|\,\|A\|_*\biggr), \quad \text{by the duality between nuclear and operator norms}\\ &\le 8\mathbb{E}(\|\Sigma_R\|)\,\gamma\sqrt{m_1m_2}\\ &\le 8C^*\sqrt{\frac{L\log d}{mn}}\gamma\sqrt{m_1m_2}, \quad \text{by Klopp (2014, Lemma 6)}\\ &\le C^*\gamma\sqrt{\frac{m_1m_2}{m}}\sqrt{\frac{L\log d}{n}} = C^*\gamma\sqrt{\frac{LM\log d}{n}},\end{aligned}$$

holds with a probability at least $1-\frac{1}{d}$. □

Lemma 2. *Suppose that T_i are i.i.d. indicator matrices with distribution Π on Γ for $i=1,\dots,n$, and Assumptions 1 and 2 hold. Then for any $A\in\mathcal{K}(\zeta,\gamma)$, the inequality*

$$\frac{1}{n}\sum_{i=1}^n\langle T_i,A\rangle^2 \ge \|A\|_{L_2(\Pi)}^2 - \zeta^2\sqrt{\frac{L\log d}{n}} - \zeta\frac{\|A\|_*}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}}$$

holds with probability at least $1-\frac{\kappa}{d}$, where κ is a constant depending on a universal constants K and L defined in Assumption 1.

Proof of Lemma 2. Let $X_i=\langle T_i,A\rangle^2$ and define

$$V := n\mathbb{E}(X_i^2) + 16n\gamma\sqrt{M}\sqrt{\frac{L\log d}{n}} \le n + 16\gamma\sqrt{LMn\log d} \lesssim n\sqrt{L\log d}. \tag{26}$$

By the Talagrand concentration inequality in Koltchinskii (2011, Theorem 2.6) and a logarithmic property that $\log(1+x)\ge\frac{x}{1+x},\forall x\ge 0$, we take $t\gtrsim\sqrt{\frac{L\log d}{n}}$, thus getting

$$\begin{aligned}\mathbb{P}\biggl\{\sup_{A\in\mathcal{K}(1,\gamma)}|Z_\gamma-\mathbb{E}(Z_\gamma)|\ge t\biggr\} &\le K\exp\biggl\{-\frac{1}{K}tn\,\log\biggl(1+\frac{tn}{n\sqrt{L\log d}}\biggr)\biggr\}\\ &\le K\exp\biggl\{-\frac{1}{K}\frac{t^2n}{\sqrt{L\log d}+t}\biggr\},\end{aligned}\tag{27}$$

where K is a universal constant, V is defined in (26), and U is the uniform bound of X_i. In our case, $U=1$.

We split the proof into two cases depending on the value of γ.

Case 1: If $\gamma\le\frac{1}{\sqrt{M}}$, then (27) implies for $t'\gtrsim 1$

$$\begin{aligned}\mathbb{P}\biggl\{\sup_{A\in\mathcal{K}(1,\gamma)}Z_\gamma\ge t'\sqrt{\frac{L\log d}{n}}\biggr\} &\le K\exp\biggl\{-\frac{1}{K}\frac{t'^2L\log d}{\sqrt{L\log d}+t'\sqrt{\frac{L\log d}{n}}}\biggr\}\\ &= K\exp\biggl\{-\frac{1}{K}\frac{t'^2\sqrt{L\log d}}{1+\frac{t'}{\sqrt{n}}}\biggr\} \le K\exp\biggl\{-\frac{1}{K}\frac{\sqrt{L\log d}}{1+\frac{t'}{\sqrt{n}}}\biggr\}.\end{aligned}\tag{28}$$

Case 2: If $\gamma\ge\frac{1}{\sqrt{M}}$, then, similarly, we have for $t'\gtrsim 1$,

$$\begin{aligned}\mathbb{P}\biggl\{\sup_{A\in\mathcal{K}(1,\gamma)}Z_\gamma\ge t'\gamma\sqrt{\frac{LM\log d}{n}}\biggr\} &\le K\exp\biggl\{-\frac{1}{K}\frac{t'^2\gamma^2LM\log d}{\sqrt{L\log d}+t'\gamma\sqrt{\frac{LM\log d}{n}}}\biggr\}\\ &= K\exp\biggl\{-\frac{1}{K}\frac{t'^2\gamma^2M\sqrt{L\log d}}{1+\gamma\sqrt{M}\frac{t'}{\sqrt{n}}}\biggr\}\\ &\le K\exp\biggl\{-\frac{1}{K}\frac{\sqrt{L\log d}}{1+\gamma\sqrt{M}\frac{t'}{\sqrt{n}}}\biggr\}.\end{aligned}\tag{29}$$

Next, we use the standard peeling argument to estimate the probability. Specifically, we define a sequence of sets

$$S_l = \left\{ A \in \mathcal{K}(1,\gamma) : 2^{l-1}\frac{1}{\sqrt{M}} \le \frac{\|A\|_*}{\sqrt{m_1m_2}} \le 2^l \frac{1}{\sqrt{M}} \right\}, \quad l = 1, 2, \cdots$$

Then a series of estimations lead to

$$\begin{aligned}
&\mathbb{P}\left\{ \sup_{\|A\|_\infty \le 1} \frac{Z_\gamma}{\|A\|_*/\sqrt{m_1m_2}} \ge t'\sqrt{\frac{LM\log d}{n}} \right\} \\
&\le \sum_{l=1}^{\infty} \mathbb{P}\left\{ \sup_{A\in S_l} \frac{Z_\gamma}{\|A\|_*/\sqrt{m_1m_2}} \ge t'\sqrt{\frac{LM\log d}{n}} \right\} \\
&\le \sum_{l=1}^{\infty} \mathbb{P}\left\{ \sup_{A\in\mathcal{K}\left(1,2^l\frac{1}{\sqrt{M}}\right)} \frac{Z_\gamma}{\|A\|_*/\sqrt{m_1m_2}} \ge t'\sqrt{\frac{LM\log d}{n}} \right\} \\
&\le \sum_{l=1}^{\infty} \mathbb{P}\left\{ \sup_{A\in\mathcal{K}\left(1,2^l\frac{1}{\sqrt{M}}\right)} Z_\gamma \ge t' 2^{l-1}\sqrt{\frac{L\log d}{n}} \right\} \\
&\le \sum_{l=1}^{\infty} K\exp\left\{ -\frac{1}{K}\frac{t'^2\, 2^{2l-2}\sqrt{L\log d}}{1+\frac{t'}{\sqrt{n}}2^{l-1}} \right\} \qquad (30) \\
&\le \sum_{l=1}^{\infty} K\exp\left\{ -\frac{1}{K}\frac{2^{l-1}\sqrt{L\log d}}{1+\frac{t'}{\sqrt{n}}} \right\}, \quad \text{because for any } l \ge 1,\ 2^{l-1} \ge 1 \\
&\le \sum_{l=1}^{\infty} K\exp\left\{ -\frac{\log 2}{2K}\frac{\sqrt{L\log d}}{1+\frac{t'}{\sqrt{n}}}\, l \right\}, \quad \text{because for any } x > 0,\ x > \log x \text{ always holds} \\
&= K\frac{\exp\left\{ -\frac{\log 2}{2K}\frac{\sqrt{L\log d}}{1+\frac{t'}{\sqrt{n}}} \right\}}{1-\exp\left\{ -\frac{\log 2}{2K}\frac{\sqrt{L\log d}}{1+\frac{t'}{\sqrt{n}}} \right\}}.
\end{aligned}$$

Based on Case 1 and Case 2, we set $t' = 1$ for large n, then for any matrix $A \in \mathbb{R}^{m_1\times m_2}$ with $\|A\|_\infty \le 1$, the following inequality

$$Z_\gamma \le \sqrt{\frac{L\log d}{n}} + \frac{\|A\|_*}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}},$$

holds with at least probability $1 - \frac{\kappa}{d}$ where κ is a constant depending on L. Therefore, we have

$$Z_\gamma \le \zeta^2\sqrt{\frac{L\log d}{n}} + \zeta\frac{\|A\|_*}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}},$$

which implies that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\langle T_i, A\rangle^2 - \|A\|^2_{L_2(\pi)}\right| \le \zeta^2\sqrt{\frac{L\log d}{n}} + \zeta\frac{\|A\|_*}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}}.$$

A simple calculation leads to the desired inequality

$$\frac{1}{n}\sum_{i=1}^{n}\langle T_i, A\rangle^2 \ge \|A\|^2_{L_2(\pi)} - \zeta^2\sqrt{\frac{L\log d}{n}} - \zeta\frac{\|A\|_*}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}},$$

which holds with probability at least $1 - \frac{\kappa}{d}$. □

Lemma 3 is important, as it gives us an upper bound on the Frobenius norm error for the estimator $\hat{A}$ under any general sampling distribution.

Lemma 3. *Suppose T_i are i.i.d. indicator matrices with distribution Π on Γ that satisfies Assumptions 1 and 2 for $i = 1, \ldots, n$. Assume $\|A_0\|_\infty \le \zeta$ for a constant ζ and Assumption 3 holds. Take $\lambda^{-1} = \mathcal{O}\left(\left(\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}\right)^{-1}\right)$, then for any $n \gtrsim d\log d$, there exist two constants C_1' and C_2' only* *depending on c_0 such that the estimator $\hat{A}$ from (5) satisfies*

$$\frac{1}{m_1m_2}\|\hat{A}-A_0\|_F^2 \le C_1'\nu\left\{(\zeta\vee\sigma)\sqrt{\frac{Ld\log d}{n}}\frac{\|A_0\|_*}{\sqrt{m_1m_2}}+\lambda\mathrm{TL1}_a(A_0)\right\}+C_2'\nu\zeta^2\sqrt{\frac{L\log d}{n}}, \tag{31}$$

with probability at least $1-\frac{\kappa+1}{d}$, where κ is a constant depending on L.

Proof of Lemma 3. It follows from the optimality of the estimator $\hat{A}$ in (5) that

$$\frac{1}{n}\sum_{i=1}^n(Y_i-\langle T_i,\hat{A}\rangle)^2+\lambda\mathrm{TL1}_a(\hat{A})\le\frac{1}{n}\sum_{i=1}^n(Y_i-\langle T_i,A_0\rangle)^2+\lambda\mathrm{TL1}_a(A_0).$$

Replacing Y_i with the trace regression model, we obtain

$$\frac{1}{n}\sum_{i=1}^n(\langle T_i,A_0\rangle+\sigma\xi_i-\langle T_i,\hat{A}\rangle)^2\le\frac{\sigma^2}{n}\sum_{i=1}^n\xi_i^2+\lambda\mathrm{TL1}_a(A_0)-\lambda\mathrm{TL1}_a(\hat{A}),$$

which, after expanding the square, is equivalent to, ,

$$\frac{1}{n}\sum_{i=1}^n\langle T_i,\hat{A}-A_0\rangle^2\le\frac{2\sigma}{n}\sum_{i=1}^n\xi_i\langle T_i,\hat{A}-A_0\rangle+\lambda\mathrm{TL1}_a(A_0)-\lambda\mathrm{TL1}_a(\hat{A}).$$

By the definition of Σ in (25), we have

$$\frac{1}{n}\sum_{i=1}^n\langle T_i,\hat{A}-A_0\rangle^2\le 2\langle\Sigma,\hat{A}-A_0\rangle+\lambda\mathrm{TL1}_a(A_0)-\lambda\mathrm{TL1}_a(\hat{A}). \tag{32}$$

We further use the duality between the nuclear norm and the operator norm to get

$$\frac{1}{n}\sum_{i=1}^n\langle T_i,\hat{A}-A_0\rangle^2\le 2\|\Sigma\|\|\hat{A}-A_0\|_*+\lambda\mathrm{TL1}_a(A_0)-\lambda\mathrm{TL1}_a(\hat{A}),$$

which indicates that

$$\lambda\mathrm{TL1}_a(\hat{A})\le 2\|\Sigma\|\|\hat{A}-A_0\|_*+\lambda\mathrm{TL1}_a(A_0). \tag{33}$$

It follows from Lemma 2 that

$$\begin{aligned}
&\frac{1}{\nu m_1m_2}\|\hat{A}-A_0\|_F^2\le\|\hat{A}-A_0\|_{L_2(\Pi)}^2\\
&\lesssim\frac{1}{n}\sum_{i=1}^n\langle T_i,\hat{A}-A_0\rangle^2+\zeta^2\sqrt{\frac{L\log d}{n}}+\zeta\frac{\|\hat{A}-A_0\|_*}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}}\\
&\lesssim 2\|\Sigma\|\|\hat{A}-A_0\|_*+\lambda\mathrm{TL1}_a(A_0)-\lambda\mathrm{TL1}_a(\hat{A})+\zeta^2\sqrt{\frac{L\log d}{n}}+\zeta\frac{\|\hat{A}-A_0\|_*}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}}\\
&\lesssim\left\{2\|\Sigma\|+\frac{\zeta}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}}\right\}\|A_0\|_*+\zeta^2\sqrt{\frac{L\log d}{n}}+\lambda\mathrm{TL1}_a(A_0)\\
&\qquad+\left\{2\|\Sigma\|+\frac{\zeta}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}}-\lambda\frac{1+a}{a+\sigma_1(\hat{A})}\right\}\|\hat{A}\|_*\\
&\lesssim\left\{2\|\Sigma\|+\frac{\zeta}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}}\right\}\|A_0\|_*+\zeta^2\sqrt{\frac{L\log d}{n}}+\lambda\mathrm{TL1}_a(A_0)\\
&\qquad+\left\{2\|\Sigma\|+\frac{\zeta}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}}-\lambda\frac{1+a}{a+\zeta\sqrt{m_1m_2}}\right\}\|\hat{A}\|_*.
\end{aligned} \tag{34}$$

The penultimate inequality holds by the triangle inequality of nuclear norm: $\|\hat{A}-A_0\|_*\le\|\hat{A}\|_*+\|A_0\|_*$ and the inequality property of $\mathrm{TL1}_a$: $\mathrm{TL1}_a(A)\ge\sum_{j=1}^m\frac{(a+1)\sigma_j(A)}{a+\sigma_1(A)}=\frac{1+a}{a+\sigma_1(A)}\|A\|_*$. In addition, since $\|\hat{A}\|_\infty\le\zeta$ and hence $\sigma_1(\hat{A})\le\zeta\sqrt{m_1m_2}$, we have the last inequality. Furthermore, using the matrix Bernstein's inequality (Klopp, 2014, Lemma 5), we have for $n\gtrsim d$ there exists a constant $C^*>0$ only depending on c_0 such that

$$\|\Sigma\|\le C^*\sigma\sqrt{\frac{L\log d}{mn}}\le C^*\frac{\sigma}{\sqrt{m_1m_2}}\sqrt{\frac{Ld\log d}{n}}, \tag{35}$$

holds with the probability at least $1 - \frac{1}{d}$.

If $\lambda^{-1} = \mathcal{O}\left(\left(\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}\right)^{-1}\right)$, then there exist two constants C_1' and C_2' depending on c_0 such that

$$\frac{1}{m_1m_2}\|\hat{A} - A_0\|_F^2 \le C_1'\nu\left\{(\zeta\vee\sigma)\sqrt{\frac{Ld\log d}{n}}\frac{\|A_0\|_*}{\sqrt{m_1m_2}} + \lambda\mathrm{TL1}_a(A_0)\right\} + C_2'\nu\zeta^2\sqrt{\frac{L\log d}{n}}, \tag{36}$$

holds with probability at least $1 - \frac{\kappa+1}{d}$, where κ is a constant depending on L. □

Equipped with Lemma 3, we are posed to prove Theorem 1 and Theorem 3.

Proof of Theorem 1. Take $\lambda \asymp \frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}$ and any $a^{-1} = \mathcal{O}((\zeta\sqrt{m_1m_2})^{-1})$, by the inequality that $\mathrm{TL1}_a(A_0) \le \frac{1+a}{a}\|A_0\|_*$, we get

$$\begin{aligned}\lambda\mathrm{TL1}_a(A_0) &\asymp \frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}\mathrm{TL1}_a(A_0)\\ &\le \frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}\frac{1+a}{a}\|A_0\|_*\\ &\lesssim (\zeta\vee\sigma)\sqrt{\frac{Ld\log d}{n}}\frac{\|A_0\|_*}{\sqrt{m_1m_2}} \le (\zeta\vee\sigma)\gamma\sqrt{\frac{Ld\log d}{n}}.\end{aligned} \tag{37}$$

It follows from Lemma 3 that there exist two constants C_1 and C_2 only depending on c_0 such that the estimator $\hat{A}$ from (5) satisfies

$$\frac{1}{m_1m_2}\|\hat{A} - A_0\|_F^2 \le C_1\nu(\zeta\vee\sigma)\gamma\sqrt{\frac{Ld\log d}{n}} + C_2\nu\zeta^2\sqrt{\frac{L\log d}{n}} \tag{38}$$

holds with probability at least $1 - \frac{\kappa+1}{d}$. □

Proof of Theorem 3. Since $\mathrm{TL1}_a$ for any $a > 0$ is an increasing function with respect to input arguments, it is straightforward that

$$\mathrm{TL1}_a(A_0) = \sum_{j=1}^{m}\frac{(1+a)\sigma_j(A_0)}{a+\sigma_j(A_0)} \le \mathrm{rank}(A_0)\frac{(1+a)\sigma_1(A_0)}{a+\sigma_1(A_0)},$$

which implies that

$$\mathrm{TL1}_a(A_0) \le \mathrm{rank}(A_0)\frac{(1+a)\zeta\sqrt{m_1m_2}}{a+\zeta\sqrt{m_1m_2}},$$

with $\sigma_1(A_0) \le \zeta\sqrt{m_1m_2}$. Applying Lemma 3 with $\lambda^{-1} = \mathcal{O}\left(\left(\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}\right)^{-1}\right)$, we obtain

$$\frac{1}{m_1m_2}\|\hat{A} - A_0\|_F^2 \lesssim \lambda\nu\frac{(1+a)\zeta\sqrt{m_1m_2}}{a+\zeta\sqrt{m_1m_2}}\mathrm{rank}(A_0) + \nu\zeta^2\sqrt{\frac{L\log d}{n}}. \tag{39}$$

Consequently, for any $a > 0$, there exist two constants C_5 and C_6 only depending on c_0 such that the estimator $\hat{A}$ from (5) satisfies

$$\frac{1}{m_1m_2}\|\hat{A} - A_0\|_F^2 \le C_5\lambda\nu\frac{(1+a)\zeta\sqrt{m_1m_2}}{a+\zeta\sqrt{m_1m_2}}\mathrm{rank}(A_0) + C_6\nu\zeta^2\sqrt{\frac{L\log d}{n}}, \tag{40}$$

with probability at least $1 - \frac{\kappa+1}{d}$. □

C Proof of Theorem 2

To prove Theorem 2, we introduce Lemma 4 along with some definitions. For any matrix $A \in \mathbb{R}^{m_1 \times m_2}$, let U_A and V_A be the left and right singular matrices of A, and D_A is the diagonal matrix with the singular values of A, i.e., the SVD of A is expressed by $A = U_A D_A V_A^\intercal$. We denote $r_A := \text{rank}(A)$ and $\sigma_j(A)$ is the jth singular values of A, $j = 1, \ldots, r_A$. We define $S_U(A)$ and $S_V(A)$ to be the linear subspaces spanned by column vectors of U_A and V_A, respectively, and denote their corresponding orthogonal components, denoted by $S_U^\perp$ and $S_V^\perp$. We set

$$P_A^\perp(B) = \mathbf{P}_{S_U^\perp(A)} B \mathbf{P}_{S_V^\perp(A)} \quad \text{and} \quad P_A(B) = B - P_A^\perp(B), \tag{41}$$

where $\mathbf{P}_S$ denotes the projection onto the linear subspace S. Then, for any matrix $B \in \mathbb{R}^{m_1 \times m_2}$, there exits the left and right singular vectors U_B, V_B and the diagonal matrix D_B with singular values of $P_A^\perp(B)$ such that $P_A^\perp(B) = U_B D_B V_B^\intercal$; similarly, let $r_B = \text{rank}(P_A^\perp(B))$, and $\sigma_j(P_A^\perp(B))$ is the jth singular values of $P_A^\perp(B)$, $j = 1, \ldots, r_B$.

Lemma 4. *For any two matrices A and B, we have*

$$\text{TL1}_a(A + P_A^\perp(B)) = \text{TL1}_a(A) + \text{TL1}_a(P_A^\perp(B)). \tag{42}$$

Proof of Lemma 4. Some calculations show that

$$A + P_A^\perp(B) = U_A D_A V_A^\intercal + U_B D_B V_B^\intercal = \begin{pmatrix} U_A & U_B \end{pmatrix} \begin{pmatrix} D_A & 0 \\ 0 & D_B \end{pmatrix} \begin{pmatrix} V_A^\intercal \\ V_B^\intercal \end{pmatrix} = UDV^\intercal, \tag{43}$$

where $U = \begin{pmatrix} U_A & U_B \end{pmatrix}$, $D = \begin{pmatrix} D_A & 0 \\ 0 & D_B \end{pmatrix}$, and $V^\intercal = \begin{pmatrix} V_A^\intercal \\ V_B^\intercal \end{pmatrix}$.

Next, we show that U is the left singular vector of $A + P_A^\perp(B)$. It is sufficient to verify that $U^\intercal U = I$, which is equivalent to $U_A^\intercal U_B = 0$. By the definition of $P_A^\perp$, it is straightforward to see that the singular vectors of $P_A^\perp(B)$ are orthogonal to the space spanned by the singular vectors of A, as SVD guarantees that U_A spans the column space of A and U_B spans the orthogonal complement of the column space of A. Therefore, $U_A^\intercal U_B = 0$ as desired. Similarly, we have $V^\intercal V = I$, implying that V is the right singular vector of $A + P_A^\perp(B)$. Hence, D is the corresponding diagonal matrix with singular values of $A + P_A^\perp(B)$.

Lastly, we apply the TL1_a function to $A + P_A^\perp(B)$, thus getting

$$\text{TL1}_a(A + P_A^\perp(B)) = \sum_{k=1}^{r_A + r_B} \frac{(1+a)D(k,k)}{a + D(k,k)} = \sum_{i=1}^{r_A} \frac{(1+a)D_A(i,i)}{a + D_A(i,i)} + \sum_{j=1}^{r_B} \frac{(1+a)D_B(j,j)}{a + D_B(j,j)} = \text{TL1}_a(A) + \text{TL1}_a(P_A^\perp(B)).$$

□

Proof of Theorem 2. We start with the derivative of $\text{TL1}_a(A)$ with respect to the matrix A. By SVD, we have $A = \sum_{i=1}^m \sigma_j(A) u_j v_j^\intercal$, which indicates that $\sigma_j(A) = u_j^\intercal A v_j$, where u_j and v_j are the left and right orthonormal singular vectors of A, respectively. Then,

$$\begin{aligned}
\frac{\partial(\text{TL1}_a(A))}{\partial A} &= \frac{\partial}{\partial A} \sum_{j=1}^m \frac{(1+a)\sigma_j(A)}{a + \sigma_j(A)} = \sum_{j=1}^m \frac{\partial}{\partial(A)} \frac{(1+a)\sigma_j(A)}{a + \sigma_j(A)} = \sum_{j=1}^m \frac{a(1+a)}{(a + \sigma_j(A))^2} \frac{\partial \sigma_j(A)}{\partial A} \\
&= \sum_{j=1}^m \frac{a(1+a)}{(a + \sigma_j(A))^2} \frac{\partial(u_j^\intercal A v_j)}{\partial A} = \sum_{j=1}^m \frac{a(1+a)}{(a + \sigma_j(A))^2} \frac{\partial(\text{tr}(A v_j u_j^\intercal))}{\partial A} = \sum_{j=1}^m \frac{a(1+a)}{(a + \sigma_j(A))^2} \frac{u_j v_j^\intercal \partial A}{\partial A} \\
&= \sum_{j=1}^m \frac{a(1+a)}{(a + \sigma_j(A))^2} u_j v_j^\intercal.
\end{aligned} \tag{44}$$

By the Mean Value Theorem, there exists a matrix $\tilde{A}$ between $\hat{A}$ and $A_0 + P_{A_0}^\perp(\hat{A} - A_0)$ such that

$$\begin{aligned}
\text{TL1}_a(\hat{A}) &= \text{TL1}_a(A_0 + \hat{A} - A_0) = \text{TL1}_a(A_0 + P_{A_0}^\perp(\hat{A} - A_0) + P_{A_0}(\hat{A} - A_0)) \\
&= \text{TL1}_a(A_0 + P_{A_0}^\perp(\hat{A} - A_0)) + \langle \nabla \text{TL1}_a(\tilde{A}), P_{A_0}(\hat{A} - A_0) \rangle \\
&\geq \text{TL1}_a(A_0) + \text{TL1}_a(P_{A_0}^\perp(\hat{A} - A_0)) - \left\| \nabla \text{TL1}_a(\tilde{A}) \right\| \| P_{A_0}(\hat{A} - A_0) \|_* \\
&\geq \text{TL1}_a(A_0) + \text{TL1}_a(P_{A_0}^\perp(\hat{A} - A_0)) - \frac{a(1+a)}{a^2} \| P_{A_0}(\hat{A} - A_0) \|_*,
\end{aligned} \tag{45}$$

where we use the penultimate inequality because of Lemma 4, the duality of the nuclear norm, and the last inequality holds because of (44).

Then, we have

$$
\begin{aligned}
&\frac{1}{\nu m_1 m_2}\|\hat{A}-A_0\|_F^2 \le \|\hat{A}-A_0\|_{L_2(\Pi)}^2 \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n}\langle T_i, \hat{A}-A_0\rangle^2 + \zeta^2\sqrt{\frac{L\log d}{n}} + \zeta\frac{\|\hat{A}-A_0\|_*}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}} \\
&\lesssim 2\|\Sigma\|\|\hat{A}-A_0\|_* + \lambda \mathrm{TL1}_a(A_0) - \lambda\mathrm{TL1}_a(\hat{A}) + \zeta^2\sqrt{\frac{L\log d}{n}} + \zeta\frac{\|\hat{A}-A_0\|_*}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}} \\
&\lesssim \left\{2\|\Sigma\| + \frac{\zeta}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}}\right\}\|\hat{A}-A_0\|_* + \lambda\mathrm{TL1}_a(A_0) - \lambda\mathrm{TL1}_a(\hat{A}) + \zeta^2\sqrt{\frac{L\log d}{n}} \\
&\lesssim \frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\sqrt{\frac{Ld\log d}{n}}\left(\|P_{A_0}^{\perp}(\hat{A}-A_0)\|_* + \|P_{A_0}(\hat{A}-A_0)\|_*\right) + \lambda\mathrm{TL1}_a(A_0) + \zeta^2\sqrt{\frac{L\log d}{n}} \\
&\qquad - \lambda\left\{\mathrm{TL1}_a(A_0) + \mathrm{TL1}_a(P_{A_0}^{\perp}(\hat{A}-A_0)) - \frac{a(1+a)}{a^2}\|P_{A_0}(\hat{A}-A_0)\|_*\right\} \\
&\lesssim \left\{\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\sqrt{\frac{Ld\log d}{n}} + \lambda\frac{a(1+a)}{a^2}\right\}\|P_{A_0}(\hat{A}-A_0)\|_* + \zeta^2\sqrt{\frac{L\log d}{n}} \\
&\qquad + \left\{\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\sqrt{\frac{Ld\log d}{n}} - \lambda\frac{1+a}{a+\zeta\sqrt{m_1m_2}}\right\}\|P_{A_0}^{\perp}(\hat{A}-A_0)\|_*,
\end{aligned}
\tag{46}
$$

where the second last inequality is obtained by (45), and the last inequality is achieved by $\mathrm{TL1}_a(A) \ge \frac{1+a}{a+\sigma_1(A)}$ and $\sigma_1(A) \le \zeta\sqrt{m_1m_2}$.

Taking $\lambda \asymp \frac{a+\zeta\sqrt{m_1m_2}}{1+a}\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\sqrt{\frac{Ld\log d}{n}}$, we get

$$
\lambda\frac{a(1+a)}{a^2} \asymp \frac{a(1+a)}{a^2}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\sqrt{\frac{Ld\log d}{n}} = \frac{a+\zeta\sqrt{m_1m_2}}{a}\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\sqrt{\frac{Ld\log d}{n}}.
$$

Together with $a^{-1} = \mathcal{O}((\zeta\sqrt{m_1m_2})^{-1})$, it deduces to

$$
\frac{1}{\nu m_1m_2}\|\hat{A}-A_0\|_F^2 \lesssim \frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\sqrt{\frac{Ld\log d}{n}}\sqrt{2\tau}\|\hat{A}-A_0\|_F + \zeta^2\sqrt{\frac{L\log d}{n}}, \tag{47}
$$

because of

$$
\|P_{A_0}(\hat{A}-A_0)\|_* \le \sqrt{\mathrm{rank}(P_{A_0}(\hat{A}-A_0))}\|\hat{A}-A_0\|_\infty \le \sqrt{2\,\mathrm{rank}(\hat{A}-A_0)}\|\hat{A}-A_0\|_F \le \sqrt{2\tau}\|\hat{A}-A_0\|_F.
$$

Therefore, we have

$$
\begin{aligned}
\frac{1}{m_1m_2}\|\hat{A}-A_0\|_F^2 &\lesssim \nu\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\sqrt{\frac{Ld\log d}{n}}\sqrt{2\tau}\|\hat{A}-A_0\|_F + \nu\zeta^2\sqrt{\frac{L\log d}{n}}, \\
\frac{1}{m_1m_2}\|\hat{A}-A_0\|_F^2 &\lesssim \nu^2(\zeta^2\vee\sigma^2)\tau\frac{Ld\log d}{n} + \nu\zeta^2\sqrt{\frac{L\log d}{n}}.
\end{aligned}
$$

In summary, there exist two constants C_3 and C_4 only depending on c_0 such that the inequality

$$
\frac{1}{m_1m_2}\|\hat{A}-A_0\|_F^2 \le C_3\nu^2(\zeta^2\vee\sigma^2)\tau\frac{Ld\log d}{n} + C_4\nu\zeta^2\sqrt{\frac{L\log d}{n}}, \tag{48}
$$

holds with probability at least $1-\frac{\kappa+1}{d}$. □

D Proof of Theorem 4 and Corollary 1

Lemma 5. *If* $\lambda^{-1} = \mathcal{O}\left(\left(\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}\right)^{-1}\right)$*, for any* $a = o((m_1m_2)^{\frac{1}{4}})$*, there exists a constant* c_1 *such that the smallest non-zero singular value of the estimator* $\hat{A} \in \mathbb{R}^{m_1\times m_2}$ *obtained by* (5) *shall be greater than or equal to* $c_1\left(a^3 + a^2\sqrt{m_1m_2}\right)^{\frac{1}{4}}$*.*

Proof of Lemma 5. Suppose the estimator matrix $\hat{A}$ is of rank k and the true matrix A_0 is of rank r_0. We ignore the oracle case when $k \le r_0$, and instead assume that $r_0 \le k \le m$. Denote the smallest non-zero singular value of $\hat{A}$ by σ_k. We prove by contradiction by assuming $\sigma_k \le c_1\left(a^3 + a^2\sqrt{m_1m_2}\right)^{\frac{1}{4}}$.

Suppose that the left and right orthonormal singular vectors of A are $\{u_j\}$ and $\{v_j\}$, while the diagonal matrix $S = \text{diag}(\sigma_1,\dots,\sigma_m)$ contains the singular values of A in decreasing order. Then we have $A = \sum_{j=1}^m \sigma_j u_j v_j^\intercal$ by SVD. Similarly, we write $A_0 = \sum_{j=1}^m \sigma_j^* u_j^* v_j^{*\intercal}$, where $\{u_j^*\}$, $\{v_j^*\}$, and $S^* = \text{diag}(\sigma_1^*,\dots,\sigma_m^*)$ are the left, right orthonormal singular vectors and the singular values of A_0, respectively. Denote $Q(A)$ to be the objective function of (5), i.e.,

$$Q(A) = \frac{1}{n}\sum_{i=1}^n\left(\langle T_i, A\rangle - Y_i\right)^2 + \lambda\text{TL1}_a(A) := l(A) + \lambda\text{TL1}_a(A),$$

where $l(A)$ denotes the least square term that can be expressed as

$$l(A) = \frac{1}{n}\sum_{i=1}^n\left(\langle T_i, \sum_{j=1}^m \sigma_j u_j v_j^\intercal\rangle - Y_i\right)^2.$$

With fixed U and V, we can compute the derivative of Q with any singular value σ_s with $s > k$

$$\begin{aligned}
\frac{\partial Q(S)}{\partial \sigma_s} &= \frac{\partial l(S)}{\partial\sigma_s} + \lambda\frac{\partial \text{TL1}_a(S)}{\partial\sigma_s} \\
&= \frac{2}{n}\sum_{i=1}^n\left[\langle T_i, \sum_{j=1}^m \sigma_j u_j v_j^\intercal\rangle - Y_i\right]\langle T_i, u_s v_s^\intercal\rangle + \lambda\frac{a(1+a)}{(a+\sigma_s)^2} \\
&= \frac{2}{n}\sum_{i=1}^n\left[\langle T_i, \sum_{j=1}^m \sigma_j^* u_j^* v_j^{*\intercal}\rangle - Y_i\right]\langle T_i, u_s v_s^\intercal\rangle \\
&\quad + \frac{2}{n}\sum_{i=1}^n\langle T_i, \sum_{j=1}^m\left(\sigma_j u_j v_j{}^\intercal - \sigma_j^* u_j^* v_j^{*\intercal}\right)\rangle\langle T_i, u_s v_s^\intercal\rangle + \lambda\frac{a(1+a)}{(a+\sigma_s)^2} \\
&= \frac{2}{n}\sum_{i=1}^n\left[\langle T_i, A_0\rangle - Y_i\right]\langle T_i, u_s v_s^\intercal\rangle + \frac{2}{n}\sum_{i=1}^n\langle T_i, \hat{A} - A_0\rangle\langle T_i, u_s v_s^\intercal\rangle + \lambda\frac{a(1+a)}{(a+\sigma_s)^2} \\
&= 2\langle\Sigma, u_s v_s^\intercal\rangle + \frac{2}{n}\sum_{i=1}^n\langle T_i, \hat{A} - A_0\rangle\langle T_i, u_s v_s^\intercal\rangle + \lambda\frac{a(1+a)}{(a+\sigma_s)^2} \\
&\le 2\|\Sigma\|\|u_s v_s^\intercal\|_* + 2\sqrt{\frac{1}{n}\sum_{i=1}^n\langle T_i, \hat{A} - A_0\rangle^2}\sqrt{\frac{1}{n}\sum_{i=1}^n\langle T_i, u_s v_s^\intercal\rangle^2} + \lambda\frac{a(1+a)}{(a+\sigma_s)^2} \\
&=: I_1 + I_2 + I_3,
\end{aligned} \tag{49}$$

where the last-second inequality is obtained by the duality of the nuclear norm and Cauchy-Schwarz inequality. We estimate the three terms I_1, I_2, and I_3 individually.

For I_1**.** Since $u_s v_s^\intercal$ is a rank-1 matrix, then $\|u_s v_s^\intercal\|_* = \|u_s\|\|v_s^\intercal\| = 1$. We further use (35) to get

$$I_1 = \mathcal{O}\left(\sqrt{\frac{Ld\log d}{nm_1m_2}}\right), \tag{50}$$

holds with probability at least $1 - \frac{1}{d}$.

For I_2**.** Following Lemma 1 and Lemma 2, we take $\gamma = 1/\sqrt{m_1 m_2}$ in the constraint set (23) and let $H_i = \langle T_i, u_s v_s^\intercal \rangle^2$ and define

$$V := n\mathbb{E}(H_i^2) + 16\frac{n}{\sqrt{m_1 m_2}}\sqrt{\frac{Ld\log d}{nm_1m_2}} = \frac{n}{m_1m_2} + 16\frac{n}{\sqrt{m_1 m_2}}\sqrt{\frac{Ld\log d}{nm_1m_2}} \lesssim \frac{n}{m_1m_2},$$

by $n \gtrsim d\log d$. Taking $t \gtrsim \sqrt{(Ld\log d)/(nm_1m_2)}$ and using the Talagrand concentration inequality (Koltchinskii, 2011, Theorem 2.6), we have for a universal constant K that

$$\mathbb{P}\Big\{\sup_{u_s v_s^\intercal \in \mathcal{K}(1,\gamma)} |Z_\gamma - \mathbb{E}(Z_\gamma)| \geq t\Big\} \leq K\exp\Big\{-\frac{1}{K}tn\ \log\left(1+tm_1m_2\right)\Big\} \leq K\exp\Big\{-\frac{1}{K}\frac{t^2 nm_1m_2}{1+tm_1m_2}\Big\}, \tag{51}$$

which implies that the inequality

$$\begin{aligned}\frac{1}{n}\langle T_i, u_s v_s^\intercal\rangle^2 &\leq \|u_s v_s^\intercal\|^2_{L_2(\Pi)} + \sqrt{\frac{Ld\log d}{nm_1m_2}} \leq \frac{\nu}{m_1m_2}\|u_s v_s^\intercal\|_F^2 + \sqrt{\frac{Ld\log d}{nm_1m_2}}\\ &= \frac{\nu}{m_1m_2} + \sqrt{\frac{Ld\log d}{nm_1m_2}} = \mathcal{O}(\sqrt{\frac{Ld\log d}{nm_1m_2}}),\end{aligned} \tag{52}$$

holds with high probability. By Lemma 3 with $\lambda^{-1} = \mathcal{O}\left(\left(\frac{(\zeta\vee\sigma)}{\sqrt{m_1m_2}}\frac{a+\zeta\sqrt{m_1m_2}}{1+a}\sqrt{\frac{Ld\log d}{n}}\right)^{-1}\right)$, we get

$$\begin{aligned}\frac{1}{n}\langle T_i, \hat{A} - A_0\rangle^2 &\leq \|\hat{A} - A_0\|^2_{L_2(\Pi)} + \zeta^2\sqrt{\frac{L\log d}{n}} + \zeta\frac{\|\hat{A}-A_0\|_*}{\sqrt{m_1m_2}}\sqrt{\frac{LM\log d}{n}}\\ &\leq \frac{\nu}{m_1m_2}\|\hat{A}-A_0\|_F^2 + \zeta^2\sqrt{\frac{L\log d}{n}} + \zeta r_0\sqrt{\frac{Ld\log d}{n}}\\ &\lesssim \nu^2\left\{(\zeta\vee\sigma)\sqrt{\frac{Ld\log d}{n}}\frac{\|A_0\|_*}{\sqrt{m_1m_2}} + \lambda\mathrm{TL1}_a(A_0)\right\} + \nu^2\zeta^2\sqrt{\frac{L\log d}{n}}\\ &= \mathcal{O}\left(\lambda(1+a)\right),\end{aligned} \tag{53}$$

where the last inequality is because of $\mathrm{TL1}_a(A_0) \leq (1+a)r_0$. Therefore, $I_2 = \mathcal{O}\left(\sqrt{\lambda(1+a)}\left(\frac{Ld\log d}{nm_1m_2}\right)^{1/4}\right)$. Together with (50), we have

$$I_1 + I_2 = \mathcal{O}\left(\sqrt{\lambda(1+a)}\left(\frac{Ld\log d}{nm_1m_2}\right)^{1/4}\right). \tag{54}$$

For I_3**.** We know that $I_3 = \lambda\frac{a(1+a)}{(a+\sigma_s)^2}$ for $s > k$. We verify whether $I_3 > I_1 + I_2$, when $a = o\left((m_1m_2)^{\frac{1}{4}}\right)$ and $\sigma_s \leq c_1\left(a^3 + a^2\sqrt{m_1m_2}\right)^{\frac{1}{4}}$. Then we have

$$\begin{aligned}\frac{I_3}{I_1+I_2} &\gtrsim \frac{a\sqrt{\lambda(1+a)}}{(a+\sigma_s)^2}\left(\frac{Ld\log d}{nm_1m_2}\right)^{-\frac{1}{4}} &(55)\\ &\gtrsim \frac{a\sqrt{1+a}}{(a+\sigma_s)^2}\sqrt{\frac{a+\zeta\sqrt{m_1m_2}}{\sqrt{m_1m_2}(1+a)}\sqrt{\frac{Ld\log d}{n}}}\left(\frac{Ld\log d}{nm_1m_2}\right)^{-\frac{1}{4}} &(56)\\ &= \frac{a\sqrt{(a+\zeta\sqrt{m_1m_2})}}{(a+\sigma_s)^2}, &(57)\end{aligned}$$

which leads to the desired bound,

$$\left(\frac{I_3}{I_1+I_2}\right)^2 \gtrsim \frac{a^2(a+\zeta\sqrt{m_1m_2})}{(a+\sigma_s)^4} > \frac{a^2(a+\zeta\sqrt{m_1m_2})}{\sigma_s^4} \gtrsim \frac{a+\zeta\sqrt{m_1m_2}}{a+\sqrt{m_1m_2}} = \mathcal{O}(1). \tag{58}$$

Overall, we have $\frac{\partial Q(S)}{\partial\sigma_s} > 0$ is always satisfied, then it follows from Fan and Li (2001, Lemma 1) that there exists a constant c_1 such that $\sigma_k \geq c_1\left(a^3 + a^2\sqrt{m_1m_2}\right)^{\frac{1}{4}}$. □

Proof of Theorem 4. By the optimality of the estimator $\hat{A}$ together with (33), we deduce

$$\lambda \mathrm{TL1}_a(\hat{A}) \leq 2\|\Sigma\|\sqrt{\mathrm{rank}(\hat{A}-A_0)}\|\hat{A}-A_0\|_F + \lambda \mathrm{TL1}_a(A_0).$$

We further use $\mathrm{TL1}_a(A_0) \leq (1+a)\mathrm{rank}(A_0)$ to get

$$\lambda \mathrm{TL1}_a(\hat{A}) \lesssim \sqrt{\frac{Ld\log d}{nm_1m_2}}\sqrt{\mathrm{rank}(\hat{A})+\mathrm{rank}(A_0)}\sqrt{m_1m_2\lambda\frac{(1+a)\zeta\sqrt{m_1m_2}}{a+\zeta\sqrt{m_1m_2}}} + \lambda(1+a)\mathrm{rank}(A_0).$$

Using $\mathrm{rank}(\hat{A}) \geq \mathrm{rank}(A_0)$, we have

$$\mathrm{TL1}_a(\hat{A}) \lesssim \lambda^{-\frac{1}{2}}\sqrt{\mathrm{rank}(\hat{A})}\sqrt{\frac{Ld\log d}{n}}\sqrt{\frac{(1+a)\zeta\sqrt{m_1m_2}}{a+\zeta\sqrt{m_1m_2}}} + (1+a)\mathrm{rank}(A_0).$$

Since $\mathrm{TL1}_a(\hat{A}) \geq \mathrm{rank}(\hat{A})\frac{(1+a)\sigma_{\mathrm{rank}(\hat{A})}(\hat{A})}{a+\sigma_{\mathrm{rank}(\hat{A})}(\hat{A})}$, we have

$$\mathrm{rank}(\hat{A}) \lesssim \left(\lambda^{-\frac{1}{2}}\sqrt{\mathrm{rank}(\hat{A})}\sqrt{\frac{Ld\log d}{n}}\sqrt{\frac{(1+a)\zeta\sqrt{m_1m_2}}{a+\zeta\sqrt{m_1m_2}}} + (1+a)\mathrm{rank}(A_0)\right)\frac{a+\sigma_{\mathrm{rank}(\hat{A})}}{(1+a)\sigma_{\mathrm{rank}(\hat{A})}},$$

which can be derived into two scenarios:

Scenario (i):

$$\mathrm{rank}(\hat{A}) \lesssim \lambda^{-1}\frac{Ld\log d}{n}\frac{\sqrt{m_1m_2}}{(1+a)(a+\sqrt{m_1m_2})}\left(\frac{a+\sigma_{\mathrm{rank}(\hat{A})}}{\sigma_{\mathrm{rank}(\hat{A})}}\right)^2; \tag{59}$$

and Scenario (ii):

$$\mathrm{rank}(\hat{A}) \lesssim \mathrm{rank}(A_0)\frac{a+\sigma_{\mathrm{rank}(\hat{A})}}{\sigma_{\mathrm{rank}(\hat{A})}}. \tag{60}$$

Using the condition that $\sigma_{\mathrm{rank}(\hat{A})}(\hat{A}) \geq c_1\left(a^3+a^2\sqrt{m_1m_2}\right)^{\frac{1}{4}}$ by Lemma 5, these two scenarios can be further simplified to

Scenario (i):

$$\begin{aligned}\mathrm{rank}(\hat{A}) &\lesssim \lambda^{-1}\frac{Ld\log d}{n}\frac{\sqrt{m_1m_2}}{(1+a)(a+\sqrt{m_1m_2})}\left(\frac{a+\left(a^3+a^2\sqrt{m_1m_2}\right)^{\frac{1}{4}}}{\left(a^3+a^2\sqrt{m_1m_2}\right)^{\frac{1}{4}}}\right)^2 \\ &\lesssim \lambda^{-1}\frac{Ld\log d}{n}\frac{\sqrt{m_1m_2}}{(1+a)(a+\sqrt{m_1m_2})}\left(\frac{\sqrt{a}}{\left(a+\sqrt{m_1m_2}\right)^{\frac{1}{4}}}+1\right)^2.\end{aligned} \tag{61}$$

Scenario (ii):

$$\mathrm{rank}(\hat{A}) \lesssim \mathrm{rank}(A_0)\frac{a+\left(a^3+a^2\sqrt{m_1m_2}\right)^{\frac{1}{4}}}{\left(a^3+a^2\sqrt{m_1m_2}\right)^{\frac{1}{4}}} = \mathrm{rank}(A_0)\left(\frac{\sqrt{a}}{\left(a+\sqrt{m_1m_2}\right)^{\frac{1}{4}}}+1\right). \tag{62}$$

Combine these two senarios (61) and (62), we have

$$\begin{aligned}\mathrm{rank}(\hat{A}) \lesssim \max\Bigg\{&\lambda^{-1}\frac{Ld\log d}{n}\frac{\sqrt{m_1m_2}}{(1+a)(a+\sqrt{m_1m_2})}\left(\frac{\sqrt{a}}{\left(a+\sqrt{m_1m_2}\right)^{1/4}}+1\right)^2, \\ &\mathrm{rank}(A_0)\left(\frac{\sqrt{a}}{\left(a+\sqrt{m_1m_2}\right)^{1/4}}+1\right)\Bigg\}.\end{aligned}$$

Hence, there exists a constant C_7 only depending on c_0 such that

$$\begin{aligned}\operatorname{rank}(\hat{A}) \leq & C_7 \left\{ \lambda^{-1} \frac{Ld \log d}{n} \frac{\sqrt{m_1 m_2}}{(1+a)(a+\sqrt{m_1 m_2})} \left(\frac{\sqrt{a}}{\left(a+\sqrt{m_1 m_2}\right)^{1/4}} + 1 \right)^2 \right. \\ & \left. + \operatorname{rank}(A_0) \left(\frac{\sqrt{a}}{\left(a+\sqrt{m_1 m_2}\right)^{1/4}} + 1 \right) \right\},\end{aligned}$$

with high probability. □

Proof of Corollary 1. Replacing λ with the order of $\frac{(\zeta \vee \sigma)}{\sqrt{m_1 m_2}} \frac{a+\zeta\sqrt{m_1 m_2}}{1+a} \sqrt{\frac{Ld \log d}{n}}$ in Theorem 3, we get

$$\begin{aligned}\frac{1}{m_1 m_2} \|\hat{A} - A_0\|_F^2 &\lesssim \nu\tau \frac{(\zeta \vee \sigma)}{\sqrt{m_1 m_2}} \frac{a+\zeta\sqrt{m_1 m_2}}{1+a} \sqrt{\frac{Ld \log d}{n}} \frac{(1+a)\zeta\sqrt{m_1 m_2}}{a+\zeta\sqrt{m_1 m_2}} + \nu\zeta^2 \sqrt{\frac{L \log d}{n}} \\ &= \nu\tau(\zeta^2 \vee \sigma^2) \sqrt{\frac{Ld \log d}{n}} + \nu\zeta^2 \sqrt{\frac{L \log d}{n}}.\end{aligned} \tag{63}$$

According to Theorem 4 regarding the upper bound on the rank of the estimator $\hat{A}$ (i.e., two scenarios) and $a = o\left((m_1 m_2)^{1/4}\right)$, we have

Scenario (i):

$$\begin{aligned}\operatorname{rank}(\hat{A}) &\lesssim \frac{(1+a)\sqrt{m_1 m_2}}{a+\sqrt{m_1 m_2}} \sqrt{\frac{n}{Ld \log d}} \frac{Ld \log d}{n} \frac{\sqrt{m_1 m_2}}{(1+a)(a+\sqrt{m_1 m_2})} \left(\frac{\sqrt{a}}{\left(a+\sqrt{m_1 m_2}\right)^{\frac{1}{4}}} + 1 \right)^2 \\ &\lesssim \frac{m_1 m_2}{(a+\sqrt{m_1 m_2})^2} \sqrt{\frac{Ld \log d}{n}} \left(\frac{\sqrt{a}}{\left(a+\sqrt{m_1 m_2}\right)^{\frac{1}{4}}} + 1 \right)^2 = \mathcal{O}_p(1),\end{aligned} \tag{64}$$

Scenario (ii):

$$\operatorname{rank}(\hat{A}) = \mathcal{O}_p(\operatorname{rank}(A_0)). \tag{65}$$

Using (64) and (65), we obtain the desired result $\operatorname{rank}(\hat{A}) = \mathcal{O}_p(\operatorname{rank}(A_0))$. □

E Parameter tuning

We compare the TL1 model (5) with nuclear-norm (Candes and Recht, 2012), max-norm (Cai and Zhou, 2016), and a hybrid approach in a combination of the nuclear norm and max norm (Fang et al., 2018). Each method has its respective hyper-parameters. We conduct a grid search to find the optimal parameters that yield the smallest relative error (RE) defined in Section 4 within the following ranges:

- For max-norm: $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$
- For hybrid:
  - $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$
  - $\mu \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$
- For nuclear norm: $\mu \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$
- For TL1:
  - $a \in \{10^{-1}, 1, 10, 20, 50, 100, 200, 500, 600, 900, 1500, 3000\}$
  - $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

All the values for λ and μ are multiplied by $\|Y\|_F$. Following Algorithm 1, we implement the TL1 model by ourselves and fix the parameter $\rho = 0.1$ as it only affects the convergence rate, but not the performance. We obtain the codes of the hybrid approach (Fang et al., 2018) from the authors, which includes the nuclear norm and the max norm as its special case. All the experiments are run on a MacBook Pro with an Apple M2 chip and 8GB of memory.

F More experimental results

Table 6 contains the results of relative errors for the reconstructed matrix to the ground truth under Scheme 1, 2, and 3 settings with noiseless and noisy data for SNR=20. Each reported value is averaged over 50 random realizations with standard deviation in parentheses. We also highlight the two best RE values in the table in a similar way as Tables 1-3.

Table 6: Similar to Tables 1-3 with SNR =20 (mean and standard deviation over 50 trials are reported)

Scheme 1 with SNR=20					
	(r, SR)	Max-norm	Hybrid	Nuclear	TL1
300	(5, 0.1)	0.327 (0.012)	*0.064 (0.013)*	0.152 (0.007)	**0.029 (0.001)**
	(5, 0.2)	0.195 (0.005)	**0.009 (0.002)**	0.068 (0.001)	*0.013 (0.001)*
	(10, 0.1)	0.568 (0.011)	*0.495 (0.013)*	0.505 (0.012)	**0.102 (0.003)**
	(10, 0.2)	0.248 (0.005)	**0.021 (0.002)**	0.120 (0.002)	*0.024 (0.001)*
500	(5, 0.1)	0.204 (0.007)	0.014 (0.000)	*0.013 (0.002)*	**0.005 (0.000)**
	(5, 0.2)	0.129 (0.004)	0.008 (0.000)	*0.007 (0.000)*	**0.003 (0.000)**
	(10, 0.1)	0.270 (0.007)	0.082 (0.006)	*0.081 (0.006)*	**0.009 (0.001)**
	(10, 0.2)	0.155 (0.003)	0.016 (0.000)	*0.011 (0.000)*	**0.006 (0.000)**

Scheme 2 with SNR=20					
	(r, SR)	Max-norm	Hybrid	Nuclear	TL1
300	(5, 0.1)	0.305 (0.033)	**0.281 (0.028)**	0.759 (0.017)	*0.368 (0.049)*
	(5, 0.2)	0.174 (0.016)	*0.139 (0.022)*	0.606 (0.020)	**0.063 (0.043)**
	(10, 0.1)	0.480 (0.026)	**0.477 (0.029)**	0.798 (0.010)	*0.505 (0.031)*
	(10, 0.2)	0.217 (0.014)	*0.204 (0.026)*	0.610 (0.016)	**0.138 (0.029)**
500	(5, 0.1)	0.265 (0.027)	*0.209 (0.022)*	0.753 (0.014)	**0.048 (0.037)**
	(5, 0.2)	0.144 (0.007)	*0.126 (0.006)*	0.606 (0.015)	**0.006 (0.000)**
	(10, 0.1)	0.337 (0.015)	*0.286 (0.016)*	0.762 (0.008)	**0.130 (0.024)**
	(10, 0.2)	0.178 (0.007)	*0.145 (0.008)*	0.609 (0.010)	**0.012 (0.011)**

Scheme 3 with SNR=20					
	(r, SR)	Max-norm	Hybrid	Nuclear	TL1
300	(5, 0.1)	0.464 (0.021)	*0.451 (0.023)*	0.784 (0.010)	**0.446 (0.031)**
	(5, 0.2)	0.184 (0.018)	*0.152 (0.024)*	0.611 (0.017)	**0.065 (0.035)**
	(10, 0.1)	0.543 (0.018)	*0.533 (0.018)*	0.820 (0.007)	**0.521 (0.018)**
	(10, 0.2)	0.229 (0.019)	*0.221 (0.022)*	0.625 (0.012)	**0.147 (0.034)**
500	(5, 0.1)	0.393 (0.025)	*0.323 (0.026)*	0.757 (0.014)	**0.107 (0.031)**
	(5, 0.2)	0.148 (0.061)	*0.127 (0.053)*	0.607 (0.014)	**0.010 (0.008)**
	(10, 0.1)	0.508 (0.012)	*0.472 (0.013)*	0.793 (0.007)	**0.329 (0.015)**
	(10, 0.2)	0.182 (0.008)	*0.151 (0.009)*	0.611 (0.010)	**0.015 (0.015)**

Additionally, we also report the performance on a larger matrix of dimension 1000×1000 across three schemes with different noise levels as well as on a matrix of dimension 1500×1500 under three schemes with $(r, SR) = (10, 0.2)$ and SNR = 20. Due to the time constraint, we can only report the results based on one random trial in Table 7 and Table 8. These results of less noisy data and larger dimensions are consistent with what we report in the main text, showing that TL1 outperforms the competing methods.

Table 7: Similar to Tables 1-3, relative errors of 1000 × 1000 matrices for one trial

Scheme 1 without noise

	(r, SR)	Max-norm	Hybrid	Nuclear	TL1
1000	(5, 0.1)	0.1135	8.87×10^{-4}	*8.77×10^{-4}*	$\mathbf{1.51 \times 10^{-5}}$
	(5, 0.2)	0.0707	*1.50×10^{-4}*	5.46×10^{-4}	$\mathbf{9.50 \times 10^{-6}}$
	(10,0.1)	0.1451	6.62×10^{-4}	1.50×10^{-3}	$\mathbf{2.60 \times 10^{-5}}$
	(10,0.2)	0.0865	*1.54×10^{-4}*	8.29×10^{-4}	$\mathbf{1.45 \times 10^{-5}}$

Scheme 1 with SNR=10

	(r, SR)	Max-norm	Hybrid	Nuclear	TL1
1000	(5, 0.1)	0.1318	0.0759	*0.0708*	**0.0696**
	(5, 0.2)	0.0811	0.0689	**0.0601**	*0.0644*
	(10,0.1)	0.1508	0.1081	*0.1043*	**0.0905**
	(10,0.2)	0.0970	0.0781	**0.0706**	*0.0754*

Scheme 2 without noise

	(r, SR)	Max-norm	Hybrid	Nuclear	TL1
1000	(5, 0.1)	0.2365	*0.1658*	0.5863	**0.0129**
	(5, 0.2)	0.0978	*0.0771*	0.6069	**0.0019**
	(10,0.1)	0.2316	*0.1599*	0.7476	**0.0716**
	(10,0.2)	0.1166	*0.0854*	0.6062	**0.0651**

Scheme 2 with SNR=10

	(r, SR)	Max-norm	Hybrid	Nuclear	TL1
1000	(5, 0.1)	0.2369	*0.1663*	0.7490	**0.0491**
	(5, 0.2)	0.0982	*0.0769*	0.6069	**0.0694**
	(10,0.1)	0.2321	*0.1609*	0.7476	**0.1091**
	(10,0.2)	0.1167	*0.0858*	0.6062	**0.0791**

Scheme 3 without noise

	(r, SR)	Max-norm	Hybrid	Nuclear	TL1
1000	(5, 0.1)	0.3064	*0.2138*	0.7490	**0.0293**
	(5, 0.2)	0.1040	*0.0768*	0.6069	**0.0001**
	(10,0.1)	0.3433	*0.2570*	0.7491	**0.1315**
	(10,0.2)	0.1162	*0.0890*	0.6062	**0.0085**

Scheme 3 with SNR=10

	(r, SR)	Max-norm	Hybrid	Nuclear	TL1
1000	(5, 0.1)	0.3064	*0.2138*	0.7490	**0.0358**
	(5, 0.2)	0.1040	*0.0768*	0.6069	**0.0133**
	(10,0.1)	0.3433	*0.2570*	0.7491	**0.1595**
	(10,0.2)	0.1162	*0.0890*	0.6062	**0.0693**

Table 8: Results, with the running time for a single trial in parentheses, for dimension 1500 × 1500 under three schemes with (r, SR) = (10, 0.2) and SNR = 20

Scheme	Max-norm	Hybrid	Nuclear	TL1
1	0.0625 (2488.8)	*0.0083* (2458.0)	0.0774 (800.64)	**0.0027** (866.07)
2	0.1048 (2364.8)	**0.0431** (2350.4)	0.5972 (753.65)	*0.0679* (798.64)
3	0.0901 (2382.6)	**0.0385** (2372.3)	0.5747 (750.75)	*0.0582* (801.10)

Table 9: The bias and variance results of the estimators derived from the TL1 and nuclear norm under Scheme 2 with SNR = 10 over 100 random trials

TL1			
$m_1 = m_2$	(r, SR)	bias^2	variance
300	(5, 0.1)	0.1666	0.0613
	(5, 0.2)	0.0028	0.0269

Nulcear norm			
$m_1 = m_2$	(r, SR)	bias^2	variance
300	(5, 0.1)	3.0098	0.0310
	(5, 0.2)	1.9715	0.0182

Table 9 shows the bias and variance results for estimators obtained from TL1 regularization and nuclear norm regularization under Scheme 2 with SNR=10 over 100 random trials.

G Real dataset description

We conduct experiments on two real datasets:

1. Coat Shopping Dataset[*] contains ratings contributed by 290 Turkers for a comprehensive inventory of 300 items. The training set comprises 6960 non-uniform, self-selected ratings, while the test set includes 4640 uniformly selected ratings. A more detailed description is provided in Schnabel et al. (2016).
2. Movielens 100K Dataset[†] is collected by the GroupLens Research Project at the University of Minnesota (Harper and Konstan, 2015). This dataset contains 100,000 ratings from 943 users across 1682 movies and is organized into ten subsets, comprising five distinct training sets, and their corresponding test sets.

[*] `https://www.cs.cornell.edu/~schnabts/mnar/`

[†] `https://www.kaggle.com/datasets/prajitdatta/movielens-100k-dataset`