Provable Low-Rank Tensor Recovery

Bo Huang*, Cun Mu*, Donald Goldfarb*, John Wright†

February 26, 2014

Abstract

In this paper, we rigorously study tractable models for provably recovering low-rank tensors. Unlike their matrix-based predecessors, current convex approaches for recovering low-rank tensors based on incomplete (tensor completion) and/or grossly corrupted (tensor robust principal analysis) observations still suffer from the lack of theoretical guarantees, although they have been used in various recent applications and have exhibited promising empirical performance. In this work, we attempt to fill this gap. Specifically, we propose a class of convex recovery models (including strongly convex programs) that can be proved to guarantee exact recovery under certain conditions. All parameters in our formulations can be determined beforehand based on the measurement data and thus there is no parameter tuning involved.

Keywords. Low-rank tensor recovery, tensor completion, tensor robust principal component analysis, Tucker decomposition, strongly convex programming, incoherence condition, sum of nuclear norms minimization.

1 Introduction

As modern computer technology keeps developing rapidly, multi-dimensional data (elements of which are addressed by more than two indices) is becoming prevalent in many areas such as computer vision [38] and information science [10, 34]. For instance, a color image is a 3-dimensional object with column, row and color modes [27]; a greyscale video is indexed by two spatial variables and one temporal variable; and 3-D face detection uses information with column, row, and depth modes. Tensor-based modeling is a natural choice in these cases because of its capability for capturing the underlying multi-linear structures. Although often residing in extremely high-dimensional spaces, the tensor of interest is frequently low-rank, or approximately so [18]. Consequently, low-rank tensor recovery or estimation is gaining significant attention in many different areas: estimating latent variable graphical models [1], classifying audio [24], mining text [7], processing radar signals [8], to name a few. Lying at the core of high-dimensional data analysis, tensor decomposition serves as a useful tool to reveal when the tensors can be modeled as lying close to a low-dimensional subspace. The two commonly used decompositions are the CANDECOMP/PARAFAC(CP) [5, 15] and Tucker

*Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027. Email: {bh2359,cm3052,goldfarb}@columbia.edu. This research was supported in part by NSF Grant DMS-1016571, ONR Grant N00014-08-1-1118, and DOE Grant DE-FG02-08ER25856.

†Department of Electrical Engineering, Columbia University, New York, NY 10027. Email: johnwright@ee.columbia.edu. John Wright is supported by Columbia University startup funding and Office of Naval Research award N00014-13-1- 0492

decomposition [37]. In particular, based on the Tucker decomposition, a convex surrogate for tensor rank, which here we refer to as the *sum-of-nuclear-norms (SNN)*, has been proposed in [23] and serves as a tractable measure of the tensor rank in practical settings.

In this work, we focus on low-rank tensor estimation under partial or corrupted observations. More specifically, we study if an underlying low-rank tensor can be recovered by minimizing the *SNN* over all tensors that obey the given data which may be incomplete or corrupted by arbitrary outliers. This idea, after first being proposed in [23], has been studied in [12, 31, 35, 36, 32], and successfully applied to various problems [33, 30, 19, 20, 11, 22]. Unlike the matrix cases, the recovery theory for low-rank tensor estimation problems is far from being well established. In [36], Tomioka et. al. conducted a statistical analysis of tensor decomposition and provided the first theoretical guarantee for *SNN* minimization. This result was significantly enhanced in a recent paper [25], in which it is not only proved that the complexity bound obtained in [36] is tight when using the *SNN* as the convex surrogate, but also proposed a simple improvement that works much better for high-order tensors. Unfortunately, both of the aforementioned results assume Gaussian measurements, while in practice the problem settings are more often similar to matrix completion [4, 28, 14] or robust PCA [3, 39] problems. Mimicking their low-dimensional predecessors, the tensor-based completion and robust PCA formulations have been applied to real applications with a promising empirical performance. However, to the best of our knowledge, it is still an open question as to what are the theoretical guarantees for exact recovery in tensor completion and tensor RPCA problems.

From the optimization perspective, efficient algorithms based on Augmented Lagrangian function or splitting techniques have been designed for low-rank tensor recovery problems, e.g., [13, 12, 40, 17]. In the matrix settings, instead of solving the original convex programming directly, some algorithms, e.g., [2, 39, 16, 9], have been proposed to solve the *strongly convex* problems by adding a small l_2 perturbation $\tau\| \cdot \|_F^2$ to the original objective. It is well known that the Lagrangian dual for a strongly convex objective is differentiable [29]. Therefore this leads to an unconstraint smooth dual problem which makes a wide class of efficient methods applicable. In [41], the L-BFGS algorithm and gradient methods with line search were studied. A gradient algorithm based on Nestrov's optimal scheme [26] is proposed in [16]. The major issue with the strongly convex approach is that for exact recovery, τ needs to tend to zero. On the other hand, empirically the convergence speed of most of the aforementioned algorithms depend on τ. In general, a larger τ leads to a faster convergence rate. Fortunately, it has been proved that a finite τ is sufficient for the purpose of low-rank matrix recovery. So a natural technical question is if the same conclusion holds for tensor completion and tensor robust PCA problems.

In this paper, we address the above two issues by providing a generalized formulation which allows for strongly convex models for both tensor completion and tensor robust PCA problems together with a rigorous study of the conditions for exact recovery for both cases. Also note that although the conclusions in this work are drawn based on the strongly convex formulation, they can be trivially generalized to their non-strictly convex cases.

1.1 Preliminary: Models for Low-rank Matrix Recovery

In this section, we review the existing convex programming models for Matrix Completion (MC) and Robust Principal Component Analysis(RPCA) [3, 39]. Both problems demonstrate that a low rank matrix $\boldsymbol{X}_0 \in \mathbb{R}^{n_1 \times n_2}$ ($n_1 \geq n_2$) can be exactly recovered from partial or/and corrupted observations via convex programming, under certain *incoherence conditions* on $\boldsymbol{X}_0$'s row and column spaces.

- **[Matrix Incoherence Conditions]** It is well known that exact recovery becomes tractable when the matrix is not in the null space of the sampling operator. This requires the singular vectors of the low-rank component $\boldsymbol{X}_0$ to be sufficiently spread and not highly correlated with any standard basis. This motivates the following definition.

  Definition 1. *Assume that $\boldsymbol{X}_0 \in \mathbb{R}^{n_1 \times n_2}$ is of rank r and has the singular value decomposition $\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, where σ_i, $1 \le i \le r$ are the singular values of $\boldsymbol{\mathcal{X}}_0$, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are the matrices of left and right singular vectors. Then the **incoherence conditions** with parameter μ are:*

  $$\max_i \left\|\boldsymbol{U}^\top \boldsymbol{e}_i\right\|^2 \le \tfrac{\mu r}{n_1}, \qquad \max_i \left\|\boldsymbol{V}^\top \boldsymbol{e}_i\right\|^2 \le \tfrac{\mu r}{n_2}, \tag{1.1}$$

  $$\|\boldsymbol{U}\boldsymbol{V}^*\|_\infty \le \sqrt{\tfrac{\mu r}{n_1 n_2}}. \tag{1.2}$$

  where $\{\boldsymbol{e}_i\}$ is the standard matrix basis.

  From (1.1) and (1.2), it follows that

  $$\max_i \|\mathcal{P}_{\boldsymbol{U}} \boldsymbol{e}_i\|_2^2 \le \frac{\mu r}{n_1} \tag{1.3}$$

  $$\max_i \|\mathcal{P}_{\boldsymbol{V}} \boldsymbol{e}_i\|_2^2 \le \frac{\mu r}{n_2} \tag{1.4}$$

  where $\mathcal{P}_{\boldsymbol{U}}(\mathcal{P}_{\boldsymbol{V}})$ is the orthogonal projection onto the column space of $\boldsymbol{U}(\boldsymbol{V})$. Thus (1.3) and (1.4) indicate how spread out the singular vectors are with respect to the standard basis. Note that for any subspace, the smallest μ can be is 1, which can be achieved when $\boldsymbol{U}$ is perfectly evenly spread out. The largest possible value for μ is n_1/r when a standard basis vector lies in the subspace spanned by the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. A well-conditioned matrix for the recovery is expected to have small incoherence parameter μ. Although other conditions such as the *rank-sparsity incoherence conditions* [6] have also been investigated, the above conditions (1.1)-(1.2) are those most commonly used for both matrix completion and RPCA problems.

- **[Matrix Completion (MC)]** In MC problems, we would like to recover the matrix $\boldsymbol{X}_0$, given that only entries is the support Ω are observed, where $\Omega \subseteq [n_1] \times [n_2]$. Namely, we observe $\mathcal{P}_\Omega[\boldsymbol{X}_0]$ where

  $$(\mathcal{P}_\Omega[\boldsymbol{X}_0])_{ij} = \begin{cases} (\boldsymbol{X}_0)_{ij}, & \text{if } (i,j) \in \Omega; \\ 0, & \text{otherwise.} \end{cases} \tag{1.5}$$

  Clearly, this problem is ill-posed for general $\boldsymbol{X}_0$. However, the low-rankness of $\boldsymbol{X}_0$ greatly alleviates the difficulty here. Here one minimizes the nuclear norm $\|\cdot\|_*$, the sum of all the singular values, to recover the original low-rank matrix. As proposed in [4], and later further studied in [28, 14], even when the number of observed entries, i.e. $|\Omega|$, is much less than the ambient dimension $n_1 n_2$, $\boldsymbol{X}_0$ with small rank r can still be exactly recovered by the following tractable (convex) approach:

  $$\begin{aligned} \min \quad & \|\boldsymbol{X}\|_* \\ \text{s.t.} \quad & \mathcal{P}_\Omega[\boldsymbol{X}] = \mathcal{P}_\Omega[\boldsymbol{X}_0]. \end{aligned} \tag{1.6}$$

  Guarantees for exactly recovering $\boldsymbol{\mathcal{X}}_0$ by solving (1.6) were first studied in [4], and later simplified and sharpened in [28, 14].

- **[Robust Principal Component Analysis (RPCA)]** In RPCA problems, the goal is to recover the low-rank matrix $\boldsymbol{X}_0$ from observations $\boldsymbol{B}$, which is a superposition of the low-rank component $\boldsymbol{X}_0$ and a sparse corruption component $\boldsymbol{E}_0$. In [3], the following convex programming problem was proposed

$$\begin{aligned} \min_{\boldsymbol{X},\boldsymbol{E}} \quad & \lambda \|\boldsymbol{X}\|_* + \|\boldsymbol{E}\|_1 \\ \text{s.t} \quad & \boldsymbol{X} + \boldsymbol{E} = \boldsymbol{B}. \end{aligned} \tag{1.7}$$

It has been shown that when $\lambda = \sqrt{n_1}$, solving (1.7) yields the exact recovery of $\boldsymbol{X}_0$ when it is low-rank and incoherent.

- **[Mixture Model]** Suppose in addition to being grossly corrupted, the data matrix $\boldsymbol{B}$ is observed only partially (say only entries in the support $\Omega \subseteq [n_1]\times[n_2]$ are accessible). The exact recovery of $\boldsymbol{X}_0$, which is considered as a combination of (1.6) and (1.7), can be accomplished by solving the following problem:

$$\begin{aligned} \min \quad & \lambda \|\boldsymbol{X}\|_* + \|\boldsymbol{E}\|_1 \\ \text{s.t.} \quad & \mathcal{P}_\Omega[\boldsymbol{X} + \boldsymbol{E}] = \boldsymbol{B}, \end{aligned} \tag{1.8}$$

where the corruption matrix $\boldsymbol{E}$ has nonzero entries only on the subset Ω of its $n_1 \times n_2$ entries, i.e., $\mathcal{P}_{\Omega^\perp}[\boldsymbol{E}] = 0$. Model (1.8) is equivalent to MC when there is no corruption, i.e., $\boldsymbol{E} = 0$, and it reduces to RPCA when Ω is the entire set of indices. This model has been studied in [3] and [21]. In particular, the bound established in [21] is consistent with the best known results for both MC and RPCA.

A strongly convex formulation is obtained by adding l_2 perturbation terms to the objective function of problem (1.8), i.e.,

$$\begin{aligned} \min \quad & \lambda \|\boldsymbol{X}\|_* + \|\boldsymbol{E}\|_1 + \frac{\tau}{2}\|\boldsymbol{X}\|_F^2 + \frac{\tau}{2}\|\boldsymbol{E}\|_F^2 \\ \text{s.t.} \quad & \mathcal{P}_\Omega[\boldsymbol{X} + \boldsymbol{E}] = \boldsymbol{B}. \end{aligned} \tag{1.9}$$

Strongly convex models have been studied for compressed sensing, MC and RPCA problems [41, 42, 43]. The results are that, instead of vanishing to zero, τ only needs to be reasonably small for exact recovery. Since an extremely small τ often leads to an unsatisfying convergence rate, this feature of τ greatly benefits optimization algorithms that utilize the strong convexity property.

1.2 Notation and Tensor Basics

Throughout the paper we denote tensors by boldface Euler script letters, e.g., $\boldsymbol{\mathcal{X}}$. Matrices are denoted by boldface capital letters, e.g., $\boldsymbol{X}$; vectors are denoted by boldface lowercase letters, e.g., $\boldsymbol{x}$; and scalars are denoted by lowercase letters, e.g., x. For the K-way tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_K}$, its mode-i *fiber* is a n_i-dimensional column vector defined by fixing every index but the ith of $\boldsymbol{\mathcal{X}}$. The mode-i *unfolding* (matricization) of the tensor $\boldsymbol{\mathcal{X}}$ is the matrix denoted by $\boldsymbol{\mathcal{X}}_{(i)} \in \mathbb{R}^{n_i\times\prod_{j\neq i} n_j}$ that is obtained by concatenating all the mode-i fibers of $\boldsymbol{\mathcal{X}}$ as column vectors. We use the notations $n_i^{(1)} = \max\{n_i, \prod_{j\neq i} n_j\}$ and $n_i^{(2)} = \min\{n_i, \prod_{j\neq i} n_j\}$. The vectorization $vec(\boldsymbol{\mathcal{X}})$ is defined as $vec(\boldsymbol{\mathcal{X}}_{(1)})$.

Tensor norms: Here we extend vector norm definitions to tensors. The Frobenius norm of any tensor $\boldsymbol{\mathcal{X}}$ is defined as

$$\|\boldsymbol{\mathcal{X}}\|_F := \|vec(\boldsymbol{\mathcal{X}})\|_2.$$

Similarly, the l_1/l_∞ norm of a tensor $\boldsymbol{\mathcal{X}}$ is defined by its vectorization, i.e.,

$$\|\boldsymbol{\mathcal{X}}\|_{1/\infty} := \|vec(\boldsymbol{\mathcal{X}})\|_{1/\infty}.$$

Tensor-matrix multiplication: The mode-i (matrix) product of a tensor $\boldsymbol{\mathcal{X}}$ with a matrix $\boldsymbol{A}$ of compatible size is denoted as $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} \times_i \boldsymbol{A}$, where the ith mode of $\boldsymbol{\mathcal{Y}}$ is

$$\boldsymbol{\mathcal{Y}}_{(i)} := \boldsymbol{A}\boldsymbol{\mathcal{X}}_{(i)}.$$

Linear and projection operators:

- [**Matricization**] We denote the tensor-to-matrix operator by a capital letter in calligraphic font, e.g., $\mathcal{A}_i$, that transforms a tensor $\boldsymbol{\mathcal{X}}$ to its mode-i unfolding, i.e.,

  $$\mathcal{A}_i(\boldsymbol{\mathcal{X}}) := \boldsymbol{\mathcal{X}}_{(i)},$$

  and the adjoint of $\mathcal{A}_i$, denoted by $\mathcal{A}_i^*$, is defined by $\mathcal{A}_i^*(\boldsymbol{\mathcal{X}}_{(i)}) = \boldsymbol{\mathcal{X}}$.

- [**Support**] For a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times \cdots n_K}$, let Ω be any subset of indices, i.e., $\Omega \in [n_1] \times [n_2] \times \cdots \times [n_K]$. Then a projection operator on $\mathcal{P}_\Omega$ is defined by

  $$(\mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}])_{i_1,i_2,\cdots i_K} := \begin{cases} \boldsymbol{\mathcal{X}}_{i_1,i_2,\cdots i_K} & (i_1, i_2, \cdots i_K) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

  The projection operator $\mathcal{P}_\Omega$ can be extended to tensor matricization. Specifically, we define the $\mathcal{P}_{\Omega_k}$ be the operator that projects the kth unfolding $\boldsymbol{\mathcal{X}}_{(k)}$ onto the support Ω, i.e.,

  $$\mathcal{P}_{\Omega_k}[\boldsymbol{\mathcal{X}}_{(k)}] := (\mathcal{A}_k \mathcal{P}_\Omega \mathcal{A}_k^*)[\boldsymbol{\mathcal{X}}_{(k)}].$$

  Also for simplicity, we denote Ω_k to be the support Ω applied to the kth mode when there is no confusions in using this notation; thus

  $$\mathcal{P}_{\Omega_k}[\boldsymbol{\mathcal{X}}_{(k)}] = (\mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}])_{(k)}. \tag{1.10}$$

Tucker decomposition: The Tucker decomposition approximates $\boldsymbol{\mathcal{X}}$ as

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_1 \boldsymbol{A}_1 \times_2 \boldsymbol{A}_2 \cdots \times_K \boldsymbol{A}_K,$$

where $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{r_1 \times r_2 \times \cdots r_K}$ is called the *core tensor*, and the factor matrices $\{\boldsymbol{A}_i \in \mathbb{R}^{n_i \times r_i}\}$ are column-wise orthonormal. The Tucker rank (also called n-rank) of $\boldsymbol{\mathcal{X}}$ is a K-dimensional vector whose i-th entry is the (matrix) rank of the mode-i unfolding $\boldsymbol{\mathcal{X}}_{(i)}$, i.e.,

$$\text{rank}_{tc}(\boldsymbol{\mathcal{X}}) := (\text{rank}(\boldsymbol{\mathcal{X}}_{(1)}), \text{rank}(\boldsymbol{\mathcal{X}}_{(2)}), \cdots \text{rank}(\boldsymbol{\mathcal{X}}_{(K)})).$$

1.3 Organization of the Paper

The paper is organized as follows. In Section 2, motivated by the non-convex and convex models widely used in practice for low-rank tensor recovery problems, we propose a class of convex recovery models (including strongly convex programs) that can be proved to guarantee exact recovery under certain conditions. Based on comparison of the empirical recovery performances between the *Sum of Nuclear Norm (SNN)* model and the *Singleton* model, we propose a new set of incoherence conditions for tensor recovery problems. In Section 3, we provide the full proof of the main theorem. The proof depends on the *Golfing scheme* that is similar to the one in [21] to construct dual certificates. We conclude the paper with a discussion about the future research direction in Section 4.

2 Low-Rank tensor recovery

Since, a tensor, generalizes the concept of a matrix, it arises naturally in applications of high-dimensional data analysis. Tensor-based low-rank recovery models including tensor completion [23] and tensor robust PCA [17] problems have been investigated and demonstrated encouraging performances in various applications. Besides the empirical studies, some progress on their theoretical guarantees have been achieved recently. In [36], Tomioka et. al. conducted a statistical study of tensor decomposition and provided the first (upper)bound on the number of random Gaussian measurements required for exact low-rank tensor recovery. More recently, Mu et. al. [25] proved that, under the same settings, the bound obtained in [36] is tight. However, both aforementioned results relate only to random Gaussian measurements, while a rigorous study for more practical settings such as tensor completion and tensor robust PCA problems has remained open. In this section, we extend the model (1.9) and propose a provable strongly convex programming model for low-rank tensor recovery problems.

- **[Convexification for tensor rank]** The most commonly used definitions for tensor rank are the CANDECOMP/PARAFAC(CP) rank [5, 15] and the Tucker rank [37]. Many recent applications focus on the Tucker rank (n-rank) because of its basis on matrix rank. Given all tensors whose corresponding elements match the given the incomplete set of observations, we would like to recovery $\boldsymbol{\mathcal{X}}_0$ by minimizing some combination of the n-vector Tucker rank, i.e.,

$$\textbf{[Completion(non-convex)]} \qquad \min_{\text{w.r.t.}\mathbb{R}_+^K} \text{rank}_{tc}(\boldsymbol{\mathcal{X}}) \qquad \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0] \tag{2.1}$$

$$\textbf{[Robust PCA(non-convex)]} \qquad \min_{\text{w.r.t.}\mathbb{R}_+^K} \text{rank}_{tc}(\boldsymbol{\mathcal{X}}) + \|\boldsymbol{\mathcal{E}}\|_0 \quad \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{B}}. \tag{2.2}$$

To convexify the NP=hard vector optimization problems (2.1)-(2.2), it is natural to replace the Tucker vector of ranks by a weighted sum of nuclear norms. This leads to the following scalar and convex optimization problems

$$\textbf{[Completion(convex)]} \qquad \min_{\boldsymbol{\mathcal{X}}} \sum_{i=1}^K \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* \qquad \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0] \tag{2.3}$$

$$\textbf{[Robust PCA(convex)]} \qquad \min_{\boldsymbol{\mathcal{X}}} \sum_{i=1}^K \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* + \|\boldsymbol{\mathcal{E}}\|_1 \quad \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{B}}. \tag{2.4}$$

The idea of using the term $\sum_{i=1}^{K} \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_*$, which we refer to as *sum-of-nuclear-norms (SNN)*, as a convex surrogate for Tucker rank was first proposed in [23]. However, in this work we assign possibly different weights to each nuclear norm, while in [23] and similar works that came afterwards, only the heuristic of an equal weighted sum of nuclear norms was considered. Consequently, more model parameters in problems (2.3) and (2.4) have to be tuned. Fortunately, we will provide an explicit expression for λ_i which allows for exact recovery that only depends on the dimensions of the targeting tensor $\boldsymbol{\mathcal{X}}$. Moreover, the above *SNN* becomes equally weighted when $\boldsymbol{\mathcal{X}}$ has the same dimension for every mode.

- **[Strongly convex programming]** From an optimization perspective, instead of directly dealing with the original convex programming problem (2.3) and (2.4), algorithms, e.g., [2, 39, 16, 9], have been proposed to solve strongly convex programming problems that approximate (2.3) and (2.4) by adding a small l_2 perturbation $\tau\|\cdot\|_F^2$ to the original objective. This has the advantage of enabling the use of faster optimization algorithms. This is due to the well known fact that the Lagrangian dual of a strongly convex program is unconstrained as well as differentiable. The main drawback of this class of problems sometimes comes from the noise introduced by the extra l_2 perturbation. The parameter τ needs to tend to zero for exact recovery. On the other hand, for most of the aforementioned algorithms, an infinitely small τ will significantly deteriorate the convergence speed. Fortunately, it has been proved that a finite τ is sufficient for the purpose of low-rank matrix recovery, and we show in this paper that the same conclusion also holds for tensor recovery problems.

2.1 Tensor Incoherence Conditions

As in low-rank matrix recovery problems, some incoherence conditions need to be met if recovery is to be possible for tensor-based problems. Hence, we propose a new set of incoherence conditions (2.5)-(2.6) for a tensor $\boldsymbol{\mathcal{X}}_0$ by extending the matrix incoherence conditions (1.1)-(1.2) to the unfoldings of $\boldsymbol{\mathcal{X}}_0$ and adding a new "mutual incoherence" condition.

Definition 2. *Suppose that for a tensor $\boldsymbol{\mathcal{X}}_0 \in \mathbb{R}^{n_1 \times n_2 \times \cdots n_K}$, its unfoldings $\{\boldsymbol{\mathcal{X}}_{(i)}\}_{i=1,2\cdots,K}$ have the singular value decompositions*

$$\boldsymbol{\mathcal{X}}_{(i)} = \boldsymbol{U}_i \boldsymbol{\Sigma}_i \boldsymbol{V}_i^\top \quad i = 1, 2, \cdots, K.$$

and let

$$\lambda_i := \sqrt{n_i^{(1)}} \quad \text{and} \quad \boldsymbol{\mathcal{T}} := \sum_{i=1}^{K} \lambda_i \mathcal{A}_i^* \boldsymbol{U}_i \boldsymbol{V}_i^\top.$$

*Then the **tensor incoherence conditions** with parameter μ are that there exists a mode k such that*

$$\textit{[k-mode incoherence]} \qquad \begin{cases} \max_j \|\boldsymbol{U}_k^\top \boldsymbol{e}_j\|^2 \le \frac{\mu r_k}{n^k}, \quad \max_j \|\boldsymbol{V}_k^\top \boldsymbol{e}_j\|^2 \le \frac{\mu r_k}{\prod_{j=1,j\neq k}^{K} n_j}, \\ \\ \|\lambda_k \boldsymbol{U}_k \boldsymbol{V}_k^\top\|_\infty \le \sqrt{\frac{\mu r_k}{n_k^{(2)}}}, \end{cases} \tag{2.5}$$

$$\textit{[mutual incoherence]} \qquad \frac{\|\boldsymbol{\mathcal{T}}\|_\infty}{K} \le \sqrt{\frac{\mu r_k}{n_k^{(2)}}}. \tag{2.6}$$

where $\{\boldsymbol{e}_i\}$ *is the standard matrix basis.*

Note that the first two inequalities in (2.5) are just the regular matrix incoherence conditions for the low-rank mode (e.g., the k-th unfolding). The second inequality of (2.5) is equivalent to

$$\|\boldsymbol{U}_k \boldsymbol{V}_k^\top\|_\infty \leq \sqrt{\frac{\mu r_k}{n_k^{(1)} n_k^{(2)}}},$$

since $\lambda_k = \sqrt{n_k^{(1)}}$. Furthermore, if we define $\kappa_i := \frac{r_i}{n_i^{(2)}}$ to be the "rank-saturation" for the ith mode, then from the triangle inequality, it follows that

$$\frac{\|\boldsymbol{\mathcal{T}}\|_\infty}{K} \leq \sqrt{\mu\kappa}, \tag{2.7}$$

where $\kappa := \max_i\{\kappa_i\}$. Obviously a large κ means that the tensor $\boldsymbol{\mathcal{X}}_0$ while having some mode (the kth) with low (Tucker)rank, also has modes with high ranks, i.e., $\boldsymbol{\mathcal{X}}_0$ is somewhat "unbalanced" with respect to its ranks. To compare the new *mutual incoherence condition* with the original matrix incoherence conditions in (2.5), we see that, if $\|\lambda_k \boldsymbol{U}_k \boldsymbol{V}_k^\top\|_\infty \leq \sqrt{\frac{\mu r_k}{n_k^{(2)}}}$ holds for some μ, then the bigger κ is, the harder it is for (2.6) to be satisfied. To demonstrate how restrictive (2.6) can be, we randomly generated a 3-way tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{100\times100\times100}$ with its Tucker decomposition

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_1 \boldsymbol{A}_1 \times_2 \boldsymbol{A}_2 \cdots \times_K \boldsymbol{A}_K,$$

where a core tensor $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{1\times r\times r}$ has entrees generated from i.i.d Gaussian distribution, and each A_i is a random orthogonal matrix. We gradually increased r from 2 to 100, so that while we always had $\kappa_1 = \frac{1}{100}$, κ ranged from $\frac{2}{100}$ to 1. From Figure 1, we observe that although r and κ were increased significantly, our mutual incoherence condition (2.6) appeared to be much looser than what (2.7) suggests, since the ratio $\frac{\|\boldsymbol{\mathcal{T}}\|_\infty/K}{\|\lambda_1 \boldsymbol{U}_1 \boldsymbol{V}_1^\top\|_\infty}$ grew much more slowly than r and κ.

Although (2.6) is not that restrictive in general, as Figure 1 illustrated, (2.6) does characterize a class of tensors on which SNN is a plausible model to use since it favors tensors whose Tucker rank is more balanced. More specifically, for problems where κ is significantly larger than κ_k so that (2.6) becomes quite restrictive, SNN minimization is more likely to perform poorly. Indeed, Figure 2 illustrates the difference between the SNN model and the Singleton model, which minimizes only the nuclear norm of the low-rank mode unfolding, on recovering an incomplete tensor $\boldsymbol{\mathcal{X}}_0$ under different ranks and levels of observations. We notice from Figure 2 that the SNN model outperforms the Singleton model when $r \leq 5$, but does worse than the Singleton model when $r > 5$. Specifically, the performance of the SNN model is very good for small r but deteriorate as r is increased, while the Singleton model usually recovers $\boldsymbol{\mathcal{X}}_0$ when the fraction of the elements of $\boldsymbol{\mathcal{X}}_0$ that are observed is greater than 0.25, regardless of the rank of all other non-low-rank modes. This is not surprising since by minimizing the sum of nuclear norms, we are enforcing a low-rank structure for all modes simultaneously even if this may not be the case for the true solution. Therefore extra conditions are needed to ensure that all the non-low-rank modes are not too far out of line from the well-conditioned low-rank mode when we are minimizing their ranks. In particular, (2.6) suggests the average overall incoherence should be on a par with the incoherence of the low-rank (kth) mode as measured by the infinity norm.

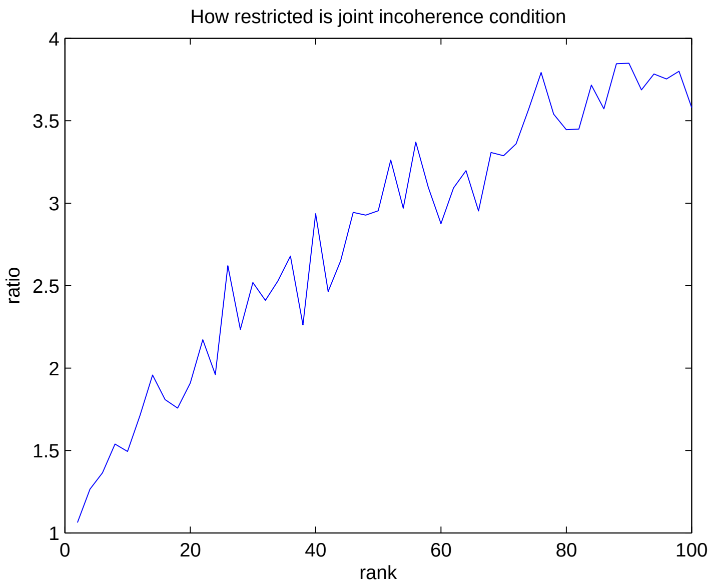

Figure 1: The average ratio $\frac{\|\sum_{i=1}^{K}\lambda_i \mathcal{A}_i^* \boldsymbol{U}_i \boldsymbol{V}_i^\top\|_\infty / K}{\|\lambda_1 \boldsymbol{U}_1 \boldsymbol{V}_1^\top\|_\infty}$ as a function of the rank r for randomly generated 3-way tensors $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{100\times100\times100}$ with Tucker rank $(1, r, r)$. For each rank $r \in [1, 100]$, we ran 10 independent trials and averaged their output ratios

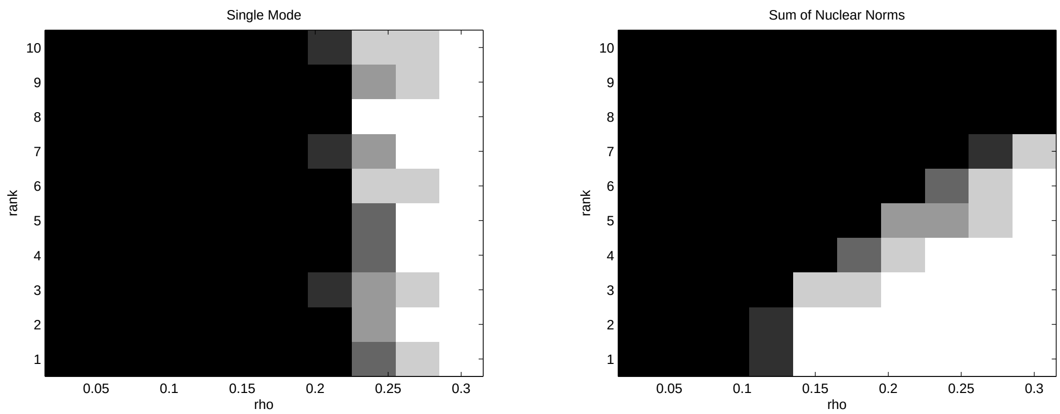

Figure 2: A random tensor $\boldsymbol{\mathcal{X}}_0 \in \mathbb{R}^{30\times30\times30}$ with (Tucker)rank $(1, r, r)$ was generated. We let r increase from 1 to 10, and the fraction ρ of the entries observed to range from 0 to 0.3. For each pair (ρ, r), we ran 5 independent trials and plotted the success rate $k/5$, where k is the number of successful recoveries, i.e., relative error $< 10^{-3}$. The lighter a region is, the more likely exact recovery can be achieved under the given choice of ρ and r.

2.2 Main Result

In this section, we consider recovering a low-rank tensor $\boldsymbol{\mathcal{X}}_0$ under incomplete and corrupted observations. Let Ω be the set of entries accessible to us. Out of the entire set Ω, a subset $\Lambda \subset \Omega$ of the entries of $\boldsymbol{\mathcal{X}}_0$ are corrupted by $\boldsymbol{\mathcal{E}}_0$, and $\Gamma \subset \Omega$ are locations where data are available and clean. As is easily seen, this can be viewed as a combination of the matrix completion and the matrix RPCA, when extended to the case of tensors.

$$\begin{aligned} \min_{\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{E}}} \quad & f(\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{E}}) := \sum_{i=1}^{K} \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* + \|\boldsymbol{\mathcal{E}}\|_1 + \frac{\tau}{2}\|\boldsymbol{\mathcal{X}}\|_F^2 + \frac{\tau}{2}\|\boldsymbol{\mathcal{E}}\|_F^2 \\ \text{s.t.} \quad & \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}}] = \boldsymbol{\mathcal{B}} \end{aligned} \tag{2.8}$$

Note that the entries of $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{B}}$ are nonzero only in support Ω, and by the definition of Γ, we also have

$$\mathcal{P}_\Gamma[\boldsymbol{\mathcal{E}}] = 0, \quad \mathcal{P}_\Gamma[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Gamma[\boldsymbol{\mathcal{B}}].$$

Our main result giving conditions under which solving (2.8) yields the exact recovery of $\boldsymbol{\mathcal{X}}_0$ is contained in the following.

Theorem 1. *Suppose $\boldsymbol{\mathcal{X}}_0$ obeys the same incoherence conditions (2.5)-(2.6) with parameters μ, and the support set Ω is uniformly distributed with cardinality $m = \rho n_k^{(1)} n_k^{(2)}$. Also suppose that each observed entry is independently corrupted with probability γ. Then provided that*

$$r_k \leq C_r K^{-2} \frac{\rho \mu^{-1} n_k^{(2)}}{\log^2 n_k^{(1)}}, \quad \gamma \leq C_\gamma \tag{2.9}$$

and

$$\tau \leq \frac{1}{2 n_k^{(1)} n_k^{(2)} (1 + \frac{4}{\rho(1-2C_\gamma)}) \|\boldsymbol{\mathcal{B}}\|_F}, \tag{2.10}$$

solving (2.8) with $\lambda_i = \sqrt{n_i^{(1)}}$ yields the exact solution $\boldsymbol{\mathcal{X}}_0$ with probability at least $1 - Cn^{-3}$ for where C, C_r and C_γ are positive numbers.

3 Architecture of the Proof

3.1 Sampling Schemes and Model Randomness

Theorem 1 is established based on using a *Uniform sampling scheme without replacement* to choose a set of entrees Ω with cardinality m. However, in order simplify our proofs, it is more convenient as is commonly done to work with other sampling schemes, such as *Bernoulli sampling*. Specifically, in order to simplify our proofs, we will work with *Bernoulli sampling* with a *Random sign model.*

[Bernoulli sampling] A *Bernoulli sampling scheme* has been used in previous work ([4], [3]) to facilitate the analysis of matrix completion and RPCA problems. For the *Bernoulli model*, we have

$\Omega := \{(i,j) : \delta_{ij} = 1\}$, where the δ_{ij}'s are i.i.d Bernoulli variables taking value one with probability ρ and zero with probability $1-\rho$. *Bernoulli sampling* can be written as $\Omega \sim \mathrm{Ber}(\rho)$ for short. Being a proxy for uniform sampling, the probability of failure under Bernoulli sampling with $p = \frac{m}{n_1 \times n_2 \cdots \times n_K}$ closely approximates the probability of failure under uniform sampling.

[Random sign model] A standard Bernoulli model assumes that

$$\begin{cases} \Lambda \sim \mathrm{Ber}(\rho\gamma) \\ \Gamma \sim \mathrm{Ber}((1-\gamma)\rho) \\ \Omega \sim \mathrm{Ber}(\rho), \end{cases}$$

and that the signs of nonzeros entries of $\boldsymbol{\mathcal{E}}_0$ are deterministic. However, it turns out that it is easier to prove Theorem 1 under the stronger assumption that the signs of the nonzeros entries of $\boldsymbol{\mathcal{E}}_0$ are independent symmetric Bernoulli variables. We define two independent random subsets of Ω:

$$\begin{aligned} \Lambda' &\sim \mathrm{Ber}(2\gamma\rho), \\ \Gamma' &\sim \mathrm{Ber}((1-2\gamma)\rho), \end{aligned}$$

It is convenient to think of

$$\boldsymbol{\mathcal{E}}_0 = \mathcal{P}_\Lambda[\boldsymbol{\mathcal{E}}],$$

for some fixed tensor $\boldsymbol{\mathcal{E}}$. Consider now a random sign tensor $\boldsymbol{\mathcal{W}}$ with i.i.d. entries such that for any index $\vec{i} \in \mathbb{R}^{i_1 \times i_2 \cdots \times i_K}$,

$$P(\boldsymbol{\mathcal{W}}_{\vec{i}} = 1) = P(\boldsymbol{\mathcal{W}}_{\vec{i}} = -1) = \frac{1}{2}.$$

Now$|\boldsymbol{\mathcal{E}}| \circ \boldsymbol{\mathcal{W}}$ has components with symmetric random signs and we define a new "noise" tensor

$$\boldsymbol{\mathcal{E}}_0' := \mathcal{P}_{\Lambda'}\left[|\boldsymbol{\mathcal{E}}| \circ \boldsymbol{\mathcal{W}}\right].$$

By the standard derandomization theory (e.g., Theorem 2.3 in [3]), *if the recovery of $\left(\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{E}}_0'\right)$ is exact with high probability, then it is also exact with at least the same probability for the model with input data $(\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{E}}_0)$*. Therefore from now on, we can equivalently work with

$$\Lambda \sim \mathrm{Ber}(2\gamma\rho), \quad \Gamma \sim \mathrm{Ber}\left((1-2\gamma)\rho\right),$$

the locations of nonzero and zero entries of $\boldsymbol{\mathcal{E}}_0$, respectively, and assume that the nonzero entries of $\boldsymbol{\mathcal{E}}_0$ have symmetric random signs.

3.2 Supporting Lemmas

Assume that the ith unfolding $\boldsymbol{\mathcal{X}}_{(i)}$ has the singular value decomposition

$$\boldsymbol{\mathcal{X}}_{(i)} = \boldsymbol{U}_i \boldsymbol{\Sigma}_i \boldsymbol{V}_i^\top \qquad i = 1, 2, \cdots, K. \tag{3.1}$$

Define T_i to be the linear space

$$T_i := \left\{ \boldsymbol{W} | \boldsymbol{W} = \boldsymbol{U}_i \boldsymbol{X}^\top + \boldsymbol{Y} \boldsymbol{V}_i^\top \quad \text{for some} \quad \boldsymbol{X}, \boldsymbol{Y} \right\}, \tag{3.2}$$

and $T_i^{\perp}$ to be the orthogonal complement of T_i. The orthogonal projection $\mathcal{P}_{T_i}$ on T_i is given by

$$\mathcal{P}_{T_i}(\boldsymbol{Z}) = \mathcal{P}_{U_i}\boldsymbol{Z} + \boldsymbol{Z}\mathcal{P}_{V_i} - \mathcal{P}_{U_i}\boldsymbol{Z}\mathcal{P}_{V_i}, \tag{3.3}$$

and $\mathcal{P}_{T_i^{\perp}}$ is defined as

$$\mathcal{P}_{T_i^{\perp}}(\boldsymbol{Z}) = (\boldsymbol{I} - \mathcal{P}_{U_i})\boldsymbol{Z}(\boldsymbol{I} - \mathcal{P}_{V_i}), \tag{3.4}$$

where $\mathcal{P}_{U_i}$ and $\mathcal{P}_{V_i}$ are the orthogonal projections onto $\boldsymbol{U}_i$ and $\boldsymbol{V}_i$ respectively.

Lemma 1. *With the tensor $\boldsymbol{\mathcal{T}}$ defined as in Definition 2, we have, for any mode i,*

$$\boldsymbol{\mathcal{T}}_{(i)} \in T_i,$$

where the subspace T_i is defined in (3.2).

Proof. For $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_1 \boldsymbol{A}_1 \times_2 \boldsymbol{A}_2 \cdots \times_K \boldsymbol{A}_K$, let $\boldsymbol{L}_k \in \mathbb{R}^{r_k \times r_k}$ and $\boldsymbol{R}_k \in \mathbb{R}^{\prod_{j\neq k} r_j \times r_k}$ be matrices of the left and right singular vectors of $\boldsymbol{\mathcal{C}}_{(k)}$, the mode k unfolding of $\boldsymbol{\mathcal{C}}$. Then the singular value decomposition of $\boldsymbol{\mathcal{X}}_{(k)}$ obeys

$$\boldsymbol{\mathcal{X}}_{(k)} = (\boldsymbol{A}_k\boldsymbol{L}_k)\,\boldsymbol{\Sigma}_k\boldsymbol{R}_k^{\top}\boldsymbol{\Phi}_k^{\top}, \tag{3.5}$$

where $\boldsymbol{\Sigma}_k \in \mathbb{R}^{r_k \times r_k}$ is the matrix whose diagonal elements are the singular values of $\boldsymbol{\mathcal{C}}_{(k)}$ and

$$\boldsymbol{\Phi}_k := \boldsymbol{A}_K \otimes \cdots \boldsymbol{A}_{k+1} \otimes \boldsymbol{A}_{k-1} \cdots \otimes A_1.$$

Therefore the subspace T_k in (3.2) corresponding to (3.5) is

$$T_k = \left\{\boldsymbol{W} | \boldsymbol{W} = \boldsymbol{A}_k\boldsymbol{L}_k\boldsymbol{X}^{\top} + \boldsymbol{Y}\boldsymbol{R}_k^{\top}\boldsymbol{\Phi}_k^{\top} \quad \text{for some} \quad \boldsymbol{X}, \boldsymbol{Y}\right\}. \tag{3.6}$$

Note that the columns of $\boldsymbol{A}_k\boldsymbol{L}_k$ are orthonormal since those of $\boldsymbol{A}_k$ are and $\boldsymbol{L}_k$ is an orthonormal matrix. On the other hand, $\boldsymbol{\mathcal{T}}$ can be explicitly written as

$$\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{C}}_T \times_1 \boldsymbol{A}_1 \times_2 \boldsymbol{A}_2 \cdots \times_K \boldsymbol{A}_K, \tag{3.7}$$

where $\boldsymbol{\mathcal{C}}_T := \left(\sum_{i=1}^{K} \lambda_i \mathcal{A}_i^{*}\boldsymbol{L}_i\boldsymbol{R}_i^{\top}\right)$, and its kth unfolding $T_{(k)} = \boldsymbol{A}_k\left(\boldsymbol{\mathcal{C}}_T\right)_{(k)}\boldsymbol{\Phi}_k$ is in T_k since we can choose $\boldsymbol{X}$ and $\boldsymbol{Y}$ in (3.6) as

$$\boldsymbol{X} = \boldsymbol{L}_k^{-1}\left(\boldsymbol{\mathcal{C}}_T\right)_{(k)}\boldsymbol{\Phi}_k, \qquad \boldsymbol{Y} = 0.$$

□

We now state three key inequalities which are crucial for the proof of the main theorem. The first and third inequalities, i.e., (3.8) and (3.10), can be found in [4] and (3.9) can be found in [3]. Note that all three inequalities are applied to the matricization on the kth mode where k is the low-rank mode.

Lemma 2. *Suppose Ω is sampled from the Bernoulli model with parameter ρ, Let $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{n_1\times n_2\cdots,\times n_K}$ and k is the low-rank mode in (2.5)-(2.6), and Ω_k is the support Ω applied to the kth mode as defined in (1.10). Then with the high probability,*

$$\left\|\rho^{-1}\mathcal{P}_{T_k}\mathcal{P}_{\Omega_k}\mathcal{P}_{T_k} - \mathcal{P}_{T_k}\right\| \leq \epsilon, \tag{3.8}$$

and

$$\left\|(\rho^{-1}P_{T_k}\mathcal{P}_{\Omega_k}\mathcal{P}_{T_k} - \mathcal{P}_{T_k})\boldsymbol{\mathcal{Z}}_{(k)}\right\|_\infty \leq \epsilon\|\mathcal{P}_{T_k}\boldsymbol{\mathcal{Z}}_{(k)}\|_\infty \tag{3.9}$$

provided that $\rho \geq C_1\epsilon^{-2}\frac{\mu r_k \log n_k^{(1)}}{n_k^{(2)}}$ for some positive numerical constant C_1;

$$\|(I - \rho^{-1}\mathcal{P}_{\Omega_k})\boldsymbol{\mathcal{Z}}_{(k)}\| \leq C_2'\sqrt{\frac{n_k^{(1)}\log n_k^{(1)}}{\rho}}\|\boldsymbol{\mathcal{Z}}_{(k)}\|_\infty \tag{3.10}$$

for some $C_2' > 0$, provided that $\rho \geq C_2\frac{\log n_k^{(1)}}{n_k^{(2)}}$ for some small constant $C_2 > 0$.

3.3 Dual Certificates

Lemma 3. *If there exists some unfolding $k \in [K]$ such that*

$$\|\frac{1}{(1-2\gamma)\rho}\mathcal{P}_{T_k}\mathcal{P}_{\Gamma_k}\mathcal{P}_{T_k} - \mathcal{P}_{T_k}\| \leq \frac{1}{2},$$

and a matrix $\boldsymbol{Y} \in \mathbb{R}^{n_k\times\prod_{j\neq k}n_j}$ satisfying

$$\begin{cases} \left\|\mathcal{P}_{T_k}[\boldsymbol{Y}] - \mathcal{P}_{T_k}\left[\boldsymbol{\mathcal{S}}_{(k)} - \boldsymbol{\mathcal{T}}_{(k)} + \tau(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}\right]\right\|_F \leq \frac{1}{n_k^{(1)}n_k^{(2)}}, \\ \left\|\mathcal{P}_{T_k^\perp}[\boldsymbol{Y}] - \mathcal{P}_{T_k^\perp}\left[\boldsymbol{\mathcal{S}}_{(k)} + \tau(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}\right]\right\| \leq \frac{\lambda_k}{2}, \\ \mathcal{P}_{\Gamma_k^\perp}[\boldsymbol{Y}] = 0, \\ \|\boldsymbol{Y}\|_\infty \leq \frac{1}{2} \end{cases} \tag{3.11}$$

where $\lambda_k = \sqrt{\rho n_k^{(1)}}$ and $\boldsymbol{\mathcal{S}}_{(k)}$ is the kth unfolding of

$$\boldsymbol{\mathcal{S}} := sgn(\boldsymbol{\mathcal{E}}_0),$$

then $(\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{E}}_0)$ is the unique solution of (2.8) when $n_k^{(1)}n_k^{(2)}$ is sufficiently large.

Proof. Consider a feasible perturbation $(\boldsymbol{\mathcal{X}}_0 + \boldsymbol{\Delta}, \boldsymbol{\mathcal{E}}_0 - \mathcal{P}_\Omega[\boldsymbol{\Delta}])$. We now show that the objective value $f(\boldsymbol{\mathcal{X}}_0 + \boldsymbol{\Delta}, \boldsymbol{\mathcal{E}}_0 - \mathcal{P}_\Omega[\boldsymbol{\Delta}])$ is strictly greater than $f(\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{E}}_0$ unless $\boldsymbol{\Delta} = \boldsymbol{0}$. Since

$$\begin{cases} \mathcal{A}_i^*[\boldsymbol{U}_i\boldsymbol{V}_i^\top + \boldsymbol{W}_i^0] & \in & \partial\left\|\mathcal{A}_i[\boldsymbol{\mathcal{X}}_0]\right\|_*, \quad \text{for any } i \in [K] \\ \boldsymbol{\mathcal{S}} + \boldsymbol{\mathcal{F}}^0 & \in & \partial\|\boldsymbol{\mathcal{E}}_0\|_1, \end{cases}$$

where for each i

$$\mathcal{P}_{T_k}[\boldsymbol{W}_i^0] = 0, \quad \left\|\boldsymbol{W}_i^0\right\| \leq 1$$
$$\mathcal{P}_{\Gamma_k^\perp}[\boldsymbol{\mathcal{F}}^0] = 0, \quad \left\|\boldsymbol{\mathcal{F}}^0\right\|_\infty \leq 1,$$

we have

$$\begin{aligned}
& f(\boldsymbol{\mathcal{X}}_0 + \boldsymbol{\Delta}, \boldsymbol{\mathcal{E}}_0 - \mathcal{P}_\Omega[\boldsymbol{\Delta}]) - f(\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{E}}_0) \\
\geq\ & \left\langle \sum_{i=1}^K \lambda_i \mathcal{A}_i^*[\boldsymbol{U}_i \boldsymbol{V}_i^*] + \sum_{i=1}^K \lambda_i \mathcal{A}_i^*[\boldsymbol{W}_i^0] + \tau \boldsymbol{\mathcal{X}}_0, \boldsymbol{\Delta} \right\rangle - \left\langle \boldsymbol{\mathcal{S}} + \boldsymbol{\mathcal{F}}^0 + \tau \boldsymbol{\mathcal{E}}_0, \mathcal{P}_\Omega[\boldsymbol{\Delta}] \right\rangle \\
=\ & \left\langle \boldsymbol{\mathcal{T}} + \sum_{i=1}^K \lambda_i \mathcal{A}_i^*[\boldsymbol{W}_i^0] + \tau \boldsymbol{\mathcal{X}}_0, \boldsymbol{\Delta} \right\rangle - \left\langle \boldsymbol{\mathcal{S}} + \boldsymbol{\mathcal{F}}^0 + \tau \boldsymbol{\mathcal{E}}_0, \boldsymbol{\Delta} \right\rangle \\
=\ & \lambda_k \left\| \mathcal{P}_{T_k^\perp}[\boldsymbol{\Delta}_{(k)}] \right\|_* + \left\| \mathcal{P}_{\Gamma_k}[\boldsymbol{\Delta}_{(k)}] \right\|_1 + \langle \boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{S}} + \tau(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0), \boldsymbol{\Delta} \rangle \\
=\ & \lambda_k \left\| \mathcal{P}_{T_k^\perp}[\boldsymbol{\Delta}_{(k)}] \right\|_* + \left\| \mathcal{P}_{\Gamma_k}[\boldsymbol{\Delta}_{(k)}] \right\|_1 + \langle \boldsymbol{Y} - \boldsymbol{\mathcal{S}}_{(k)} + \boldsymbol{\mathcal{T}}_{(k)} + \tau(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}, \boldsymbol{\Delta}_{(k)} \rangle - \langle \boldsymbol{Y}, \boldsymbol{\Delta}_{(k)} \rangle \\
=\ & \lambda_k \left\| \mathcal{P}_{T_k^\perp}[\boldsymbol{\Delta}_{(k)}] \right\|_* + \left\| \mathcal{P}_{\Gamma_k}[\boldsymbol{\Delta}_{(k)}] \right\|_1 + \langle \mathcal{P}_{T_k}\left[\boldsymbol{Y} - \boldsymbol{\mathcal{S}}_{(k)} + \boldsymbol{\mathcal{T}}_{(k)} + \tau(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}\right], \mathcal{P}_{T_k}\left[\boldsymbol{\Delta}_{(k)}\right] \rangle \\
& + \langle \mathcal{P}_{T_k^\perp}\left[\boldsymbol{Y} - \boldsymbol{\mathcal{S}}_{(k)} + \tau(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}\right], \mathcal{P}_{T_k^\perp}\left[\boldsymbol{\Delta}_{(k)}\right] \rangle + \langle \boldsymbol{Y}, \mathcal{P}_{\Gamma_k}\left[\boldsymbol{\Delta}_{(k)}\right] \rangle \\
\geq\ & \frac{\lambda_k}{2} \left\| \mathcal{P}_{T_k^\perp}[\boldsymbol{\Delta}_{(k)}] \right\|_* + \frac{1}{2} \left\| \mathcal{P}_{\Gamma_k}[\boldsymbol{\Delta}_{(k)}] \right\|_1 - \frac{1}{n_k^{(1)} n_k^{(2)}} \left\| \mathcal{P}_{T_k}[\boldsymbol{\Delta}_{(k)}] \right\|_F,
\end{aligned} \tag{3.12}$$

where the first inequality follows directly from the the convexity of $\|\cdot\|_*$ and $\|\cdot\|_1$; the second inequality holds as

$$\mathcal{P}_{\Omega^\perp}\left[\boldsymbol{\mathcal{S}} + \boldsymbol{\mathcal{F}}^0 + \tau \boldsymbol{\mathcal{E}}_0\right] = 0;$$

The third equality requires choosing $\boldsymbol{W}_i^0 = 0$ for all $i \neq k$ and picking up $\boldsymbol{W}_k^0$ and $\boldsymbol{\mathcal{F}}^0$ such that

$$\begin{aligned}
\langle \mathcal{A}_k^* \boldsymbol{W}_k^0, \boldsymbol{\Delta} \rangle &= \langle \boldsymbol{W}_k^0, \boldsymbol{\Delta}_{(k)} \rangle = \|\mathcal{P}_{T_k^\perp}[\boldsymbol{\Delta}_{(k)}]\|_* \\
\langle \boldsymbol{\mathcal{F}}^0, \Delta \rangle &= \|\mathcal{P}_\Gamma[\boldsymbol{\Delta}]\|_1 = \|\mathcal{P}_{\Gamma_k}[\boldsymbol{\Delta}_{(k)}]\|_1;
\end{aligned}$$

the last inequality is due to (3.11), thus

$$\begin{aligned}
\langle \mathcal{P}_{T_k}\left[\boldsymbol{Y} - \boldsymbol{\mathcal{S}}_{(k)} + \boldsymbol{\mathcal{T}}_{(k)} + \tau(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}\right], \mathcal{P}_{T_k}\left[\boldsymbol{\Delta}_{(k)}\right] \rangle &\geq -\frac{1}{n_k^{(1)} n_k^{(2)}} \left\| \mathcal{P}_{T_k}[\boldsymbol{\Delta}_{(k)}] \right\|_F \\
\langle \mathcal{P}_{T_k^\perp}\left[\boldsymbol{Y} - \boldsymbol{\mathcal{S}}_{(k)} + \tau(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}\right], \mathcal{P}_{T_k^\perp}\left[\boldsymbol{\Delta}_{(k)}\right] \rangle &\geq -\frac{\lambda_k}{2} \left\| \mathcal{P}_{T_k^\perp}[\boldsymbol{\Delta}_{(k)}] \right\|_* \\
\langle \boldsymbol{Y}, \mathcal{P}_{\Gamma_k}\left[\boldsymbol{\Delta}_{(k)}\right] \rangle &\geq -\frac{1}{2} \left\| \mathcal{P}_{\Gamma_k}[\boldsymbol{\Delta}_{(k)}] \right\|_1
\end{aligned}$$

Recall that we have

$$\left\| \frac{1}{(1-2\gamma)\rho} \mathcal{P}_{T_k} \mathcal{P}_{\Gamma_k} \mathcal{P}_{T_k} - \mathcal{P}_{T_k} \right\| \leq \frac{1}{2},$$

which implies $\|\frac{1}{\sqrt{(1-2\gamma)\rho}}\mathcal{P}_{T_k}\mathcal{P}_{\Gamma_k}\| \leq \sqrt{3/2}$, then

$$\begin{aligned}
\|\mathcal{P}_{T_k}[\boldsymbol{\Delta}_{(k)}]\|_F &\leq 2\left\|\frac{1}{(1-2\gamma)\rho}\mathcal{P}_{T_k}\mathcal{P}_{\Gamma_k}\mathcal{P}_{T_k}[\boldsymbol{\Delta}_{(k)}]\right\|_F \\
&\leq 2\left\|\frac{1}{(1-2\gamma)\rho}\mathcal{P}_{T_k}\mathcal{P}_{\Gamma_k}\mathcal{P}_{T_k^\perp}[\boldsymbol{\Delta}_{(k)}]\right\|_F + 2\left\|\frac{1}{(1-2\gamma)\rho}\mathcal{P}_{T_k}\mathcal{P}_{\Gamma_k}[\boldsymbol{\Delta}_{(k)}]\right\|_F \\
&\leq \sqrt{\frac{6}{(1-2\gamma)\rho}}\left\|\mathcal{P}_{T_k^\perp}[\boldsymbol{\Delta}_{(k)}]\right\|_F + \sqrt{\frac{6}{(1-2\gamma)\rho}}\left\|\mathcal{P}_{\Gamma_k}[\boldsymbol{\Delta}_{(k)}]\right\|_F \qquad (3.13)
\end{aligned}$$

Substituting (3.13) into (3.12), we obtain

$$\begin{aligned}
& f(\boldsymbol{\mathcal{X}}_0+\boldsymbol{\Delta}, \boldsymbol{\mathcal{E}}_0-\mathcal{P}_\Omega[\boldsymbol{\Delta}]) - f(\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{E}}_0) \\
\geq\; & \left(\frac{\lambda_k}{2} - \frac{1}{n_k^{(1)}n_k^{(2)}}\sqrt{\frac{6}{(1-2\gamma)\rho}}\right)\left\|\mathcal{P}_{T_k}[\boldsymbol{\Delta}_{(k)}]\right\|_F \\
& + \left(\frac{1}{2} - \frac{1}{n_k^{(1)}n_k^{(2)}}\sqrt{\frac{6}{(1-2\gamma)\rho}}\right)\left\|\mathcal{P}_{\Gamma_k}[\boldsymbol{\Delta}_{(k)}]\right\|_F. \qquad (3.14)
\end{aligned}$$

When $n_k^{(1)}n_k^{(2)}$ is large such that

$$\frac{1}{2} - \frac{1}{n_k^{(1)}n_k^{(2)}}\sqrt{\frac{6}{(1-2\gamma)\rho}} > 0,$$

the inequality (3.14) holds if and only if $\mathcal{P}_{T_k}[\boldsymbol{\Delta}_{(k)}] = \mathcal{P}_{\Gamma_k}[\boldsymbol{\Delta}_{(k)}] = 0$. On the other hand, when ρ is small such that

$$\|\mathcal{P}_{T_k}\mathcal{P}_{\Gamma_k}\| \leq \sqrt{\frac{3(1-2\gamma)\rho}{2}} < 1,$$

which implies that $\mathcal{P}_{T_k}\mathcal{P}_{\Gamma_k}$ is injective. As a result, (3.14) holds if and only if $\boldsymbol{\Delta} = 0$. □

Proof of Theorem 1:

Proof. We apply the *Golfing Scheme* similar to that in [21] to construct the dual certificate $\boldsymbol{Y}$ that satisfies

$$\begin{cases}
\left\|\mathcal{P}_{T_k}\boldsymbol{Y} - \mathcal{P}_{T_k}\left[\boldsymbol{\mathcal{S}}_{(k)} - \boldsymbol{\mathcal{T}}_{(k)}\right]\right\|_F \leq \frac{1}{2n_k^{(1)}n_k^{(2)}} \\
\left\|\mathcal{P}_{T_k^\perp}\boldsymbol{Y}\right\| \leq \frac{\lambda_k}{8} \\
\left\|\mathcal{P}_{T_k^\perp}\left[\boldsymbol{\mathcal{S}}_{(k)}\right]\right\| \leq \frac{\lambda_k}{8} \\
\|\boldsymbol{Y}\|_\infty \leq \frac{1}{2}
\end{cases} \qquad (3.15)$$

and verify the following condition for τ

$$\begin{cases}
\tau \cdot \left\|\mathcal{P}_{T_k}\left[(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}\right]\right\|_F \leq \frac{1}{2n_k^{(1)}n_k^{(2)}} \\
\tau \cdot \left\|\mathcal{P}_{T_k^\perp}\left[(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}\right]\right\| \leq \frac{\lambda_k}{4}.
\end{cases} \qquad (3.16)$$

.

[**Proof of** (3.15)] We construct $\boldsymbol{Y}$, which is supported on Γ_k, by gradually increasing the size of Γ_k. Now think of $\Gamma_k \sim \text{Ber}((1-2\gamma)\rho)$ as a union of sets of support Γ^j, i.e., $\Gamma_k = \bigcup_{j=1}^{p} \Gamma^j$ where $\Gamma^j \sim \text{Ber}(q_j)$. Define $q_1 = q_2 = \frac{(1-2\gamma)\rho}{6}$ and $q_3 = \cdots = q_p = q$, which implies $q \geq C\rho/\log n_k^{(1)}$. Thus we have

$$1-(1-2\gamma)\rho = \left(1 - \frac{(1-2\gamma)\rho}{6}\right)^2 (1-q)^{p-2},$$

where $p = \lfloor 5\log n + 1 \rfloor$. Starting from $\boldsymbol{Y}_0 = 0$, we define $\boldsymbol{Y}^L$ inductively

$$\begin{cases} \boldsymbol{Z}^0 = \mathcal{P}_{T_k}\left[\boldsymbol{\mathcal{S}}_{(k)} - \boldsymbol{\mathcal{T}}_{(k)}\right], \\ \boldsymbol{Y}^j = \sum_{i=1}^{j} q_i^{-1} \mathcal{P}_{\Gamma^i}[\boldsymbol{Z}^{i-1}], \\ \boldsymbol{Z}^j = \boldsymbol{Z}^0 - \boldsymbol{\mathcal{P}_{T_k}} Y^j, \end{cases}$$

which implies that

$$\boldsymbol{Z}^j = \left(\mathcal{P}_{T_k} - q_j^{-1}\mathcal{P}_{T_k}\mathcal{P}_{\Gamma^j}\mathcal{P}_{T_k}\right)[\boldsymbol{Z}^{j-1}].$$

Then it follows from Lemma 2 that

$$\begin{aligned} &\|\boldsymbol{Z}^j\|_F \leq \frac{1}{2}\|\boldsymbol{Z}^{j-1}\|_F \\ &\|\boldsymbol{Z}^1\|_\infty \leq \frac{1}{2\sqrt{\log n_k^{(1)}}}\|\boldsymbol{Z}^0\|_\infty, \quad \|\boldsymbol{Z}^j\|_\infty \leq \frac{1}{2^j \log n_k^{(1)}}\|\boldsymbol{Z}^0\|_\infty \quad \forall j > 1, \\ &\left\|\left(I - q_j^{-1}\mathcal{P}_{\Gamma^j}\right)\boldsymbol{Z}^{j-1}\right\| \leq C\sqrt{\frac{n_k^{(1)}\log n_k^{(1)}}{q_j}}\|\boldsymbol{Z}^{j-1}\|_\infty \end{aligned}$$

- We first bound $\left\|\boldsymbol{Z}^0\right\|_F$ and $\left\|\boldsymbol{Z}^0\right\|_\infty$. By the triangle inequality, we have

$$\|\boldsymbol{Z}^0\|_\infty \leq \|\boldsymbol{\mathcal{T}}_{(k)}\|_\infty + \|\mathcal{P}_{T_k}[\boldsymbol{\mathcal{S}}_{(k)}]\|_\infty.$$

Recall that for any $(i,j) \in \mathbb{R}^{n_k^{(1)} \times n_k^{(2)}}$, we have

$$\|\mathcal{P}_{T_k}[e_i e_j^\top]\|_\infty \leq \frac{2\mu r_k}{n_k^{(2)}}, \quad \|\mathcal{P}_{T_k}[e_i e_j^\top]\|_F \leq \sqrt{\frac{2\mu r_k}{n_k^{(2)}}}.$$

By Bernstein's inequality, we have

$$\begin{aligned} &\mathbb{P}\left(|\langle \mathcal{P}_{T_k}\left[\boldsymbol{\mathcal{S}}_{(k)}\right], e_i e_j^\top\rangle| \geq t\right) \\ = \ &\mathbb{P}\left(|\langle \boldsymbol{\mathcal{S}}_{(k)}, \mathcal{P}_{T_k}[e_i e_j^\top]\rangle| \geq t\right) \\ \leq \ &2\exp\left(-\frac{t^2/2}{N + Mt/3}\right), \end{aligned}$$

where

$$N := 2\gamma\rho \cdot \|\mathcal{P}_{T_k} e_i e_j^\top\|_F^2 \leq C\gamma\rho\frac{\mu r_k}{n_k^{(2)}},$$

and

$$M := \left\|\mathcal{P}_{T_k} e_i e_j^\top\right\|_\infty \leq \frac{2\mu r_k}{n_k^{(2)}}.$$

Then with high probability, we have

$$\|\mathcal{P}_{T_k}[\boldsymbol{\mathcal{S}}_{(k)}]\|_\infty \leq C\sqrt{\rho\frac{\mu r_k \log n_k^{(1)}}{n_k^{(2)}}},$$

and from the mutual incoherence condition (2.6)

$$\|\boldsymbol{\mathcal{T}}_{(k)}\|_\infty = \|\boldsymbol{\mathcal{T}}\|_\infty \leq K\sqrt{\frac{\mu r_k}{n_k^{(2)}}}$$

Therefore we have

$$\|\boldsymbol{Z}^0\|_\infty \leq CK\sqrt{\frac{\mu r_k \log n_k^{(1)}}{n_k^{(2)}}} \tag{3.17}$$

$$\|\boldsymbol{Z}^0\|_F \leq \sqrt{n_k^{(1)} n_k^{(2)}}\|\boldsymbol{Z}^0\|_\infty \leq CK\sqrt{\mu r_k n_k^{(1)} \log n_k^{(1)}} \tag{3.18}$$

- Second, we bound $\|\mathcal{P}_{T_k^\perp}\boldsymbol{Y}^p\|$.

$$\begin{aligned}
\|\mathcal{P}_{T_k^\perp}\boldsymbol{Y}^p\| &\leq \sum_j \|q_j^{-1}\mathcal{P}_{T_k^\perp}\mathcal{P}_{\Gamma^j}\boldsymbol{Z}^{j-1}\| \\
&\leq \sum_j \|q_j^{-1}\left(\mathcal{P}_{\Gamma^j} - I\right)\boldsymbol{Z}^{j-1}\| \\
&\leq C\sum_j \sqrt{\frac{n_k^{(1)}\log n_k^{(1)}}{q_j}}\|\boldsymbol{Z}^{j-1}\|_\infty \\
&\leq C\sqrt{n_k^{(1)}\log n_k^{(1)}}\left(\sum_{j=3}^p \frac{1}{2^{j-1}\log n_k^{(1)}\sqrt{q_j}} + \frac{1}{2\sqrt{\log n_k^{(1)}}q_2} + \frac{1}{\sqrt{q_1}}\right)\|\boldsymbol{Z}^0\|_\infty \\
&\leq CK\sqrt{\frac{n_k^{(1)}\mu r_k\left(\log n_k^{(1)}\right)^2}{n_k^{(2)}\rho}} \\
&\leq C\sqrt{\frac{n_k^{(1)}}{C_\rho}} \\
&\leq \frac{\lambda_k}{8}.
\end{aligned}$$

The last inequality holds when C_ρ is large enough.

- Third, we bound $\left\|\mathcal{P}_{T_k^\perp}[\boldsymbol{\mathcal{S}}_{(k)}]\right\|$. Since $\left\|\mathcal{P}_{T_k^\perp}[\boldsymbol{\mathcal{S}}_{(k)}]\right\| \leq \|\boldsymbol{\mathcal{S}}_{(k)}\|$ and the sign matrix $\boldsymbol{\mathcal{S}}_{(k)} = \mathrm{sgn}(\boldsymbol{\mathcal{E}}_0)$ is distributed as

$$\left(\boldsymbol{\mathcal{S}}_{(k)}\right)_{ij} = \begin{cases} 1, & w.p. \quad \gamma\rho \\ 0, & w.p. \quad 1-2\gamma\rho \\ -1, & w.p. \quad \gamma\rho, \end{cases}$$

standard arguments about the norm of a matrix with i.i.d entries give

$$\|\boldsymbol{\mathcal{S}}_{(k)}\| \leq \frac{\lambda_k}{8},$$

when C_γ is sufficiently small.

- Fourth, we bound $\|\boldsymbol{Y}^p\|_\infty$.

$$\begin{aligned}
\|\boldsymbol{Y}^p\|_\infty &\leq \sum_j \left\|q_j^{-1}\mathcal{P}_{T_k^\perp}\mathcal{P}_{\Gamma^j}\boldsymbol{Z}^{j-1}\right\|_\infty \\
&\leq \sum_j \left\|q_j^{-1}\left(\mathcal{P}_{\Gamma^j} - I\right)\boldsymbol{Z}^{j-1}\right\|_\infty \\
&\leq C\sum_j \frac{1}{q_j}\|\boldsymbol{Z}^{j-1}\|_\infty \\
&\leq \left(\sum_{j=3}^{p} \frac{1}{2^{j-1}\log n_k^{(1)}\sqrt{q_j}} + \frac{1}{2\sqrt{\log n_k^{(1)} q_2}} + \frac{1}{\sqrt{q_1}}\right)\|\boldsymbol{Z}^0\|_\infty \\
&\leq K\sqrt{\frac{\mu r \log n_k^{(1)}}{n_k^{(2)}\rho}} \\
&\leq \sqrt{\frac{1}{C_\rho \log n_k^{(1)}}} \\
&\leq 1/4
\end{aligned}$$

provided C_ρ is sufficiently large.

- Last, we show that

$$\left\|\mathcal{P}_{T_k}\boldsymbol{Y}^p - \mathcal{P}_{T_k}\left[\boldsymbol{\mathcal{S}}_{(k)} - \boldsymbol{\mathcal{T}}_{(k)}\right]\right\|_F \leq \frac{1}{2n_k^{(1)}n_k^{(2)}}.$$

Since $\mathcal{P}_{T_k}\boldsymbol{Y}^p - \mathcal{P}_{T_k}\left[\boldsymbol{\mathcal{S}}_{(k)} - \boldsymbol{\mathcal{T}}_{(k)}\right] = \mathcal{P}_{T_k}\boldsymbol{Y}^p - \boldsymbol{Z}^0 = -\boldsymbol{Z}^p$, we only need to bound $\|\boldsymbol{Z}^p\|_F$, i.e.,

$$\begin{aligned}
\left\|\mathcal{P}_{T_k}\boldsymbol{Y}^p - \mathcal{P}_{T_k}\left[\boldsymbol{\mathcal{S}}_{(k)} - \boldsymbol{\mathcal{T}}_{(k)}\right]\right\|_F &= \|\boldsymbol{Z}^p\|_F \\
&\leq C\left(\frac{1}{2}\right)^p \|\boldsymbol{Z}^0\|_F \\
&\leq C\left(n_k^{(1)}\right)^{-5}\sqrt{\mu r_k n_k^{(1)}\log n_k^{(1)}} \\
&\leq \frac{1}{2n_k^{(1)}n_k^{(2)}}.
\end{aligned}$$

[Proof of (3.16)] To establish the condition for τ under which (3.16) holds, it suffices to bound $\|\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0\|_F$ since

$$\left\|\mathcal{P}_{T_k}\left[(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}\right]\right\|_F \le \|\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0\|_F,$$
$$\left\|\mathcal{P}_{T_k^\perp}\left[(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0)_{(k)}\right]\right\| \le \|\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0\|_F,$$

and we require

$$\tau < \frac{1}{2n_k^{(1)}n_k^{(2)}\|\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0\|_F}.$$

We observe that

$$\|\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0\|_F = \|(\mathcal{I} + \mathcal{P}_\Omega)[\boldsymbol{\mathcal{X}}_0] - \boldsymbol{\mathcal{B}}\|_F \le 2\|\boldsymbol{\mathcal{X}}_0\|_F + \|\boldsymbol{\mathcal{B}}\|_F. \tag{3.19}$$

Since

$$\|\mathcal{P}_{T_k} - \frac{1}{(1-2\gamma)\rho}\mathcal{P}_{T_k}\mathcal{P}_{\Gamma_k}\mathcal{P}_{T_k}\| \le \frac{1}{2},$$

we have

$$\|\boldsymbol{\mathcal{X}}_0\|_F \le \frac{2}{(1-2\gamma)\rho}\left\|\mathcal{P}_{T_k}\mathcal{P}_{\Gamma_k}\mathcal{P}_{T_k}[(\boldsymbol{\mathcal{X}}_0)_{(k)}]\right\|_F = \frac{2}{(1-2\gamma)\rho}\left\|\mathcal{P}_{T_k}\mathcal{P}_{\Gamma_k}[\boldsymbol{\mathcal{B}}_{(k)}]\right\|_F \le \frac{2}{(1-2\gamma)\rho}\|\boldsymbol{\mathcal{B}}\|_F. \tag{3.20}$$

Therefore we have

$$\|\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0\|_F \le \left(1 + \frac{4}{(1-2\gamma)\rho}\right)\|\boldsymbol{\mathcal{B}}\|_F,$$

and it suffices to have

$$\tau \le \frac{1}{2n_k^{(1)}n_k^{(2)}\left(1 + \frac{4}{p_0(1-\gamma_s)}\right)\|\boldsymbol{\mathcal{B}}\|_F}, \tag{3.21}$$

□

4 Discussions and Conclusions

In this paper, we establish a theoretical bound for the low-rank tensor recovery model, i.e., (2.8), based on SNN convexification. The model (2.8), extending both matrix completion and matrix RPCA models to the case of tensors, employs a strongly convex programming formulation. Its reduced form, i.e., $\tau = 0$, has been repeatedly used in practice with a promising empirical performance. This is, to the best of our knowledge, the first rigorous study on theoretical guarantees for the aforementioned problems. Simulations suggest that the tensor model (2.8) is not superior to the matrix model unconditionally, thus we propose a new set of tensor incoherence condition under which using high-order tensor models based on SNN minimization is plausible.

Although we establish a set of sufficient conditions for exactly recovering a low-rank tensor, it remains unclear how to derive necessary conditions. In [25], the authors provided, when minimizing the SNN, a necessary condition that $\Omega\left(rn^{k-1}\right)$ Gaussian measurements are required to exactly recover a k-way tensor with dimension n and Tucker rank r on each mode. It would be more interesting to extend such arguments to the more practical tensor completion and tensor RPCA models discussed in this work. Furthermore, the bound achieved in Theorem 1 does not explain why the SNN model is superior to the Singleton model, as suggested by numerical simulations, when the Tucker ranks of all modes are simultaneously low. This requires a sharper bound for the SNN model, which is an unaddressed issue in this work and serves as an interesting topic for future research.

References

[1] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *CoRR abs/1210.7559* (2012).

[2] Cai, J.-F., Candès, E. J., and Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization 20*, 4 (2010), 1956–1982.

[3] Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *arXiv preprint arXiv:0912.3599* (2009).

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

[5] Carroll, J. D., and Chang, J.-J. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika 35*, 3 (1970), 283–319.

[6] Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization 21*, 2 (2011), 572–596.

[7] Cohen, S. B., and Collins, M. Tensor decomposition for fast latent-variable PCFG parsing. In *Proceedings of NIPS* (2012).

[8] D. Nion, N. S. Tensor algebra and multi-dimensional harmonic retrieval in signal processing for mimo radar. *IEEE Trans. on Signal Processing 58*, 11 (2010), 5693–5705.

[9] Fountoulakis, K., and Gondzio, J. A second-order method for strongly convex l1-regularization problems. *arXiv preprint arXiv:1306.5386* (2013).

[10] Franz, T., Schultz, A., Sizov, S., and Staab, S. Triplerank: Ranking semantic web data by tensor decomposition. In *The Semantic Web-ISWC 2009*. Springer, 2009, pp. 213–228.

[11] G. Ely, S. Aeron, N. H., and Kilmer, M. E. 5d and 4d pre-stack seismic data completion using tensor nuclear norm (tnn). *preprint* (2013).

[12] Gandy, S., Recht, B., and Yamada, I. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems 27*, 2 (2011), 025010.

[13] Goldfarb, D., and Qin, Z. Robust low-rank tensor recovery: Models and algorithms. *preprint* (2012).

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

[15] Harshman, R. A. Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis.

[16] Huang, B., Ma, S., and Goldfarb, D. Accelerated linearized bregman method. *Journal of Scientific Computing 54*, 2-3 (2013), 428–453.

[17] Jiang, B., Ma, S., and Zhang, S. Tensor principal component analysis via convex optimization. *arXiv preprint arXiv:1212.2702* (2012).

[18] Kolda, T. G., and Bader, B. W. Tensor decompositions and applications. *SIAM Rev. 51*, 3 (2009), 455–500.

[19] Kreimer, N., S. A., and Sacchi, M. D. Nuclear norm minimization and tensor completion in exploration seismology. In *International Conference on Acoustics, Speech and Signal Processing* (2013).

[20] Li, N., and Li, B. Tensor completion for on-board compression of hyperspectral images. In *Image Processing (ICIP), 2010 17th IEEE International Conference on* (2010), pp. 517–520.

[21] Li, X. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation 37*, 1 (2013), 73–99.

[22] Li, Y., Yan, J., Zhou, Y., and Yang, J. Optimum subspace learning and error correction for tensors. In *ECCV* (2010), pp. 790–803.

[23] Liu, J., Musialski, P., Wonka, P., and Ye, J. Tensor completion for estimating missing values in visual data. In *ICCV* (2009), pp. 2114–2121.

[24] Mesgarani, N., Slaney, M., and Shamma, S. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Trans. Audio, Speech, and Language Processing 14*, 3 (2006), 920–930.

[25] Mu, C., Huang, B., Wright, J., and Goldfarb, D. Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870* (2013).

[26] Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical Programming 103*, 1 (2005), 127–152.

[27] Plataniotis, K. N., and Venetsanopoulos, A. N. *Color image processing and applications*. Springer, 2000.

[28] Recht, B. A simpler approach to matrix completion. *arXiv preprint arXiv:0910.0651* (2009).

[29] Rockafellar, R. T. *Convex analysis*, vol. 28. Princeton university press, 1997.

[30] Semerci, O., Hao, N., Kilmer, M., and Miller, E. Tensor based formulation and nuclear norm regularizatin for multienergy computed tomography. *IEEE Transactions on Image Processing* (2013).

[31] Signoretto, M., De Lathauwer, L., and Suykens, J. A. Nuclear norms for tensors and their use for convex multilinear estimation. *Submitted to Linear Algebra and Its Applications 43* (2010).

[32] Signoretto, M., Tran Dinh, Q., Lathauwer, L., and Suykens, J. Learning with tensors: a framework based on convex optimization and spectral regularization. *Machine Learning* (2013), 1–49.

[33] Signoretto, M., Van de Plas, R., De Moor, B., and Suykens, J. A. K. Tensor versus matrix completion: A comparison with application to spectral data. *IEEE SPL 18*, 7 (2011), 403–406.

[34] Sun, J.-T., Zeng, H.-J., Liu, H., Lu, Y., and Chen, Z. Cubesvd: a novel approach to personalized web search. In *Proceedings of the 14th international conference on World Wide Web* (2005), ACM, pp. 382–390.

[35] Tomioka, R., Hayashi, K., and Kashima, H. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789* (2010).

[36] Tomioka, R., Suzuki, T., Hayashi, K., and Kashima, H. Statistical performance of convex tensor decomposition. *Advances in Neural Information Processing Systems (NIPS)* (2011), 137.

[37] Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika 31*, 3 (1966), 279–311.

[38] Vasilescu, M. A. O., and Terzopoulos, D. Multilinear subspace analysis of image ensembles. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on* (2003), vol. 2, IEEE, pp. II–93.

[39] Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in neural information processing systems 22* (2009), 2080–2088.

[40] Yang, L., Huang, Z.-H., and Shi, X. A fixed point iterative method for low -rank tensor pursuit. *Signal Processing, IEEE Transactions on 61*, 11 (2013), 2952–2962.

[41] Yin, W. Analysis and generalizations of the linearized bregman method. *SIAM Journal on Imaging Sciences 3*, 4 (2010), 856–877.

[42] You, Q., Wan, Q., and Liu, Y. Strongly convex programming for principal component pursuit. *arXiv preprint arXiv:1209.4405* (2012).

[43] Zhang, H., Cai, J.-F., Cheng, L., and Zhu, J. Strongly convex programming for exact matrix completion and robust principal component analysis. *arXiv preprint arXiv:1112.3946* (2011).