Provable Models for Robust Low-Rank Tensor Completion

Bo Huang[1], Cun Mu[1], Donald Goldfarb[1], John Wright[2]

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 a set of new tensor incoherence conditions which only require the existence of one low-rank mode, and characterize the problems where our models tend to perform well.

Keywords: robust low-rank tensor completion, tensor robust principal component analysis, Tucker decomposition, strongly convex programming, incoherence conditions, sum of nuclear norms.
Mathematics Subject Classification: 15A69, 90C25, 47N10, 90C59

---

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 [41] and information science [10, 37]. For instance, a color image is a 3-dimensional object with column, row and color modes [30]; 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 these underlying multi-linear structures. Although often residing in extremely high-dimensional spaces, the tensor of interest is frequently of low-rank, or approximately so [19]. Consequently, low-rank tensor recovery or estimation is gaining significant attention in many different areas: estimating latent variable graphical models [1], classifying audio [26], 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 for revealing when a tensor can be modeled as lying close to a low-dimensional subspace. The two commonly used decompositions are the CANDECOMP/PARAFAC(CP) [6, 15] and Tucker decomposition [40]. 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 [25] and has since appeared frequently in practical settings.

In this work, we focus on the Robust Low-rank Tensor Completion (RLRTC) problem which recovers a low-rank tensor from 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 [25], has been studied in [12, 34, 38, 39, 35], and successfully applied to various problems [36, 33, 20, 22, 11, 24]. Unlike the matrix cases, the recovery theory for low-rank tensor estimation problems is far from being well established. In [39], Tomioka et. al. conducted a statistical analysis of tensor decomposition and provided the first theoretical guarantee for *SNN* minimization. This result was further enhanced by [27], in which it is not only proved that the sample complexity bound obtained in [39] is tight when using the *SNN* as the convex surrogate, but also an alternative model that works much better for high-order tensors is proposed. Unfortunately, both of the aforementioned results assumed Gaussian measurements, while in practice the problem settings are more often similar to matrix completion [5, 31, 14] or robust PCA [4, 42] problems. Mimicking their low-dimensional predecessors, the RLRTC models that minimize the SNN have been applied to real applications exhibiting promising empirical performances. However, to the best of our knowledge, there are still open

---

[1]Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027. Email: {bh2359,cm3052,goldfarb}@columbia.edu. BH, CM and DG were supported by NSF Grant DMS-1016571. CM was also supported by the Class of 1988 Doctoral Fellowship.

[2]Department of Electrical Engineering, Columbia University, New York, NY 10027. Email: johnwright@ee.columbia.edu. JW was supported by Columbia University startup funding and Office of Naval Research award N00014-13-1-0492.

questions regarding the theoretical guarantees for exact recovery in SNN-based tensor completion and tensor RPCA problems.

From an optimization perspective, efficient algorithms based on the Augmented Lagrangian method and splitting techniques have been designed for RLRTC problems, e.g., [13, 12, 44, 18]. In the matrix setting, instead of solving the original convex programming problem directly, some algorithms, e.g., [3, 42, 16, 9], have been proposed to solve the *strongly convex* problems obtained 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 [32]. Therefore this leads to an unconstrained smooth dual problem which makes a wide class of efficient methods applicable. In [45], the L-BFGS algorithm and gradient methods with line search were studied, and a gradient algorithm based on Nesterov's optimal scheme [28] was 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. Therefore a natural technical question is if the same conclusion holds for tensor recovery problems.

In this paper, we conduct a thorough study of models for RLRTC problems. Our contributions are three-fold.

1. We provide provable models for RLRTC problems based on the *Sum of Nuclear Norms* (SNN) convexification of the Tucker rank of a tensor. We also show that under mild conditions, the SNN model can achieve exact recovery by automatically identifying the low-rank mode, and moreover, prior knowledge of which mode is the one that is low-rank mode is not required. This is somewhat surprising, and it significantly broadens the applicability of the model in practice, especially when it is difficult to identify the exact nature of the low-rank property of the target tensor.

2. We compared the empirical performance of the SNN model against the *Singleton* matrix model which only minimizes the nuclear norm of the lowest rank mode. Our experiments show that there exists different regimes in which the high-order SNN model is superior to the regular matrix model and vice versa. Inspired by these results, we derive a new set of *tensor incoherence conditions* for exact recovery, which enables characterizing when the SNN model is a plausible model to use.

3. Due to the stronger convergence properties of fast algorithms that utilize strong convexity, we enhance our models with a parameterized term that makes them strongly convex. We further show that the strong convexity parameter τ does not need to tend to zero for exact recovery, thus facilitating the use of these much more efficient algorithms.

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) [4, 42]. 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 1.1. 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 $\sigma_i,\ 1 \leq i \leq r$ 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:

$$\max_i \ \left\|\boldsymbol{U}^\top \boldsymbol{e}_i\right\|_2^2 \leq \tfrac{\mu r}{n_1}, \qquad \max_i \ \left\|\boldsymbol{V}^\top \boldsymbol{e}_i\right\|_2^2 \leq \tfrac{\mu r}{n_2}, \tag{1.1}$$

$$\left\|\boldsymbol{U}\boldsymbol{V}^\top\right\|_\infty \leq \sqrt{\tfrac{\mu r}{n_1 n_2}}, \tag{1.2}$$

  where $\{\boldsymbol{e}_i\}$ is the standard matrix basis and $\|\boldsymbol{U}\boldsymbol{V}^\top\|_\infty$ is the ℓ_∞ norm of the vectorization of $\boldsymbol{U}\boldsymbol{V}^\top$.

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

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

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

  where $\mathcal{P}_{\boldsymbol{U}}$ (resp. $\mathcal{P}_{\boldsymbol{V}}$) is the orthogonal projection onto the column space of $\boldsymbol{U}$ (resp. $\boldsymbol{V}$). 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. To the best of our knowledge, the above conditions (1.1)-(1.2) are those most commonly used for analyzing the exact recovery 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 in the support Ω are observed, where $\Omega \subseteq [n_1] \times [n_2]$. Namely, we observe $\mathcal{P}_\Omega[\boldsymbol{X}_0]$ where

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

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 [5], and later further studied in [31, 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:

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

Guarantees for exactly recovering $\boldsymbol{\mathcal{X}}_0$ by solving (1.6) were first studied in [5], and later simplified and sharpened in [31, 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 [4], the following convex programming problem was proposed

$$\begin{array}{ll} \min\limits_{\boldsymbol{X},\boldsymbol{E}} & \lambda \|\boldsymbol{X}\|_* + \|\boldsymbol{E}\|_1 \\ \text{s.t.} & \boldsymbol{X} + \boldsymbol{E} = \boldsymbol{B}. \end{array} \tag{1.7}$$

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

- **[Consolidated 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:

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

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 [4] and [23]. In particular, the bound established in [23] 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.,

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

Strongly convex models have been studied for compressed sensing, MC and RPCA problems [45, 46, 48]. 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 notation $n_i^{(1)} = \max\{n_i, \prod_{j\neq i} n_j\}$ and $n_i^{(2)} = \min\{n_i, \prod_{j\neq i} n_j\}$. The vectorization $vec(\boldsymbol{\mathcal{X}})$ is defined as $vec(\boldsymbol{\mathcal{X}}_{(1)})$.

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 the l_1/l_∞ norm of its vectorization, i.e.,

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

We use the norm $\|\cdot\|$ with no subscript to denote the spectral norm for matrices and Euclidean norm for vectors.

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

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

  The projection operator $\mathcal{P}_\Omega$ can be extended to tensor matricization. Specifically, we define $\mathcal{P}_{\Omega_k}$ to 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 the support Ω applied to the kth mode to be Ω_k when there is no confusions in using this notation.

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* with $r_i \ll n_i$ in the low-rank case, 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.,

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

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) for which we can 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 somewhat similar to the one in [23] to construct dual certificates. We conclude the paper with a discussion about future research directions in Section 4.

2 Robust Low-Rank Tensor Completion

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 [25] and tensor robust PCA [13] problems have been investigated and shown to perform well in various applications. Besides these empirical studies, some progress on their theoretical guarantees have been achieved recently. In [39], 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. [27] proved that, under the same settings, the bound obtained in [39] is tight. However, both aforementioned results relate only to random Gaussian measurements, while a rigorous study for more practical settings such as RLRTC problems, which are sometimes also known as tensor completion and tensor robust PCA problems, has remained open. In this section, we extend the model (1.9) and propose a strongly convex programming model for RLRTC problems with a provable recovery guarantee.

- **[Convexification for tensor rank]** The most commonly used definitions for tensor rank are the CANDECOMP/PARAFAC(CP) rank [6, 15] and the Tucker rank [40]. Many recent applications focus on the Tucker rank (n-rank) because of its basis on matrix rank. Given all tensors whose corresponding elements match an incomplete set of observations, we would like to recover $\boldsymbol{\mathcal{X}}_0$ by minimizing some combination of the n-vector Tucker rank, i.e.,

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

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

$$\textbf{[Mixture (non-convex)]} \quad \min_{\text{w.r.t.}\mathbb{R}_+^K} \operatorname{rank}_{tc}(\boldsymbol{\mathcal{X}}) + \|\boldsymbol{\mathcal{E}}\|_0 \quad \text{s.t. } \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}}] = \boldsymbol{\mathcal{B}}. \tag{2.3}$$

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

$$\textbf{[Completion (convex)]} \quad \min_{\boldsymbol{\mathcal{X}}} \sum_{i=1}^K \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* \quad \text{s.t. } \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0]; \tag{2.4}$$

$$\textbf{[Robust PCA (convex)]} \quad \min_{\boldsymbol{\mathcal{X}}} \sum_{i=1}^K \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* + \|\boldsymbol{\mathcal{E}}\|_1 \quad \text{s.t. } \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{B}}; \tag{2.5}$$

$$\textbf{[Consolidated (convex)]} \quad \min_{\boldsymbol{\mathcal{X}}} \sum_{i=1}^K \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* + \|\boldsymbol{\mathcal{E}}\|_1 \quad \text{s.t. } \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}}] = \boldsymbol{\mathcal{B}}. \tag{2.6}$$

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 the Tucker rank was first proposed in [25]. However, in this work we assign possibly different weights to each nuclear norm, while in [25] 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 (2.4) and (2.5) have to be tuned. Fortunately, we can provide an explicit expression for λ_i that allows for exact recovery and only depends on the dimensions of the target tensor $\boldsymbol{\mathcal{X}}$. Moreover, the above SNN becomes equally weighted when $\boldsymbol{\mathcal{X}}$ has the same dimension for every mode.

- **[Comparison with the matrix model]** Suppose that we have some prior knowledge about the rank of the underlying tensor, i.e., the kth mode is low rank. Inspired by the matrix recovery model, we can recover the true tensor by minimizing the nuclear norm of only the kth mode, which leads to the following Singleton model

$$\min_{\boldsymbol{\mathcal{X}}_{(k)}} \|\boldsymbol{\mathcal{X}}_{(k)}\|_* + \|\boldsymbol{\mathcal{E}}_{(k)}\|_1 \qquad \text{s.t. } \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_{(k)} + \boldsymbol{\mathcal{E}}_{(k)}] = \boldsymbol{B}_{(k)}. \tag{2.7}$$

Note that (2.7) is exactly (1.8), with k being the low-rank mode and $\boldsymbol{\mathcal{E}}_{(k)}$ and $\boldsymbol{\mathcal{B}}_{(k)}$ being the kth unfoldings of the corruption and observation tensors, respectively. Although the recovery theory for the Singleton model has been well studied, it cannot be directly applied to the non-degenerate SNN model (2.6), where $\lambda_i \neq 0$ for all $1 \leq i \leq K$. In fact, the empirical performance of these two models can be rather different and the success of one does not necessarily imply the success of the other. To demonstrate this, we randomly generated a 3-way tensor $\boldsymbol{\mathcal{X}}_0 \in \mathbb{R}^{30\times30\times30}$ with Tucker rank $(1, r, r)$.

We let r increase from 1 to 10, and the fraction ρ of the entries observed to range from 0 to 0.3. We set equal weights for the SNN term as in [25] and similar works that came afterwards, i.e.,

$$\lambda_i = 1, \qquad i = 1, 2 \cdots K.$$

For simplicity, we assume no corruption existed in this case, i.e.

$$\boldsymbol{\mathcal{E}} = \mathbf{0}.$$

Therefore (2.6) reduces to the completion problem (2.4), and Figure 1 illustrates the success rates of the two models.[1]

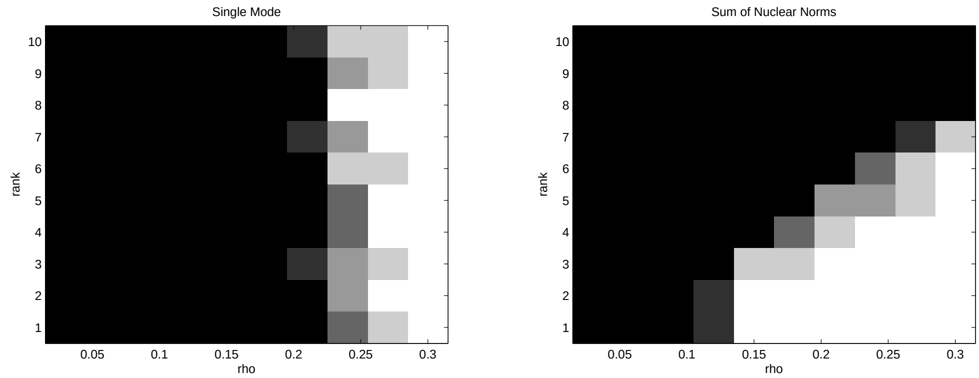

Figure 1: A random tensor $\boldsymbol{\mathcal{X}}_0 \in \mathbb{R}^{30\times 30\times 30}$ with Tucker rank $(1, r, r)$ was generated. We let r increase from 1 to 10, and the fraction ρ of the entries observed to range from 0 to 0.3. For each pair (ρ, r), we ran 5 independent trials and plotted the 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.

We notice from Figure 1 that the equal weighted SNN model outperformed the Singleton model when $r \leq 5$, but did worse than the Singleton model when $r > 5$. For instance, when $r = 1$, the SNN model succeeds using only half of the number of observations of the Singleton model. When r increases to 10, the SNN model fails while the Singleton model is capable of exact recovery with about 25% observations. Specifically, the performance of the SNN model was very good for small r but deteriorated as r was increased, while the Singleton model usually recovered $\boldsymbol{\mathcal{X}}_0$ when the fraction of the elements of $\boldsymbol{\mathcal{X}}_0$ that were observed was greater than 0.25, regardless of the rank of all of the 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 the next section we derive a new tensor *mutual incoherence condition* that characterizes the regimes where the SNN model tends to work better for RLRTC problems.

- **[Strongly convex programming]** From an optimization perspective, instead of directly dealing with the original convex programming problem (2.4) and (2.5), algorithms, e.g., [3, 42, 16, 9], have been proposed to solve strongly convex programming problems that approximate (2.4) and (2.5) 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

[1] Note that many efficient algorithms have been proposed for solving the above Singleton and SNN models, e.g., [3, 42, 16, 9, 13, 12, 44, 18]. Here we applied the Accelerated Linearizd Bregman (ALB) algorithm proposed in [16] to solve both models. For the Singleton model, more implementation details can be found in [16]. Solving SNN model with ALB has been demonstrated in the Appendix of [27]

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 arguments can be extended to RLRTC 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 (2.8)-(2.9) for the sought after 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. In contrast to the prevailing assumptions on low-rank tensors, which require all modes to have low rank, our tensor incoherence conditions indicate that the existence of only one low-rank mode is sufficient for exact recovery, and we do not need to know in advance which mode is of low rank.

Definition 2.1. 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

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

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

$$\textbf{[k-mode incoherence]} \quad \begin{cases} \max_j \|\boldsymbol{U}_k^\top \boldsymbol{e}_j\|_2^2 \le \frac{\mu r_k}{n^k}, \quad \max_j \|\boldsymbol{V}_k^\top \boldsymbol{e}_j\|_2^2 \le \frac{\mu r_k}{\prod_{j=1, j\neq k}^{K} n_j}, \\ \|\lambda_k \boldsymbol{U}_k \boldsymbol{V}_k^\top\|_\infty \le \mu \sqrt{\frac{r_k}{n_k^{(2)}}}, \end{cases} \tag{2.8}$$

$$\textbf{[mutual incoherence]} \quad \frac{\|\boldsymbol{\mathcal{T}}\|_\infty}{K} \le \mu \sqrt{\frac{r_k}{n_k^{(2)}}}. \tag{2.9}$$

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

Note that the first two inequalities in (2.8) are just the regular matrix incoherence conditions for the low-rank mode (e.g., the k-th unfolding). If we define $\mu_k := \mu$ where μ is defined in (2.8) and (2.9), the second inequality of (2.8) is equivalent to

$$\|\boldsymbol{U}_k \boldsymbol{V}_k^\top\|_\infty \le \mu_k \sqrt{\frac{r_k}{n_k^{(1)} n_k^{(2)}}}, \tag{2.10}$$

since $\lambda_k = \sqrt{n_k^{(1)}}$. Furthermore, in order to account for the effect of the other modes and compare with the condition (2.9), we generalize the above relation (2.10) to all modes, i.e. an incoherence parameter μ_i is chosen for mode i such that the following inequality holds,

$$\|\boldsymbol{U}_i \boldsymbol{V}_i^\top\|_\infty \le \mu_i \sqrt{\frac{r_i}{n_i^{(1)} n_i^{(2)}}} \quad i = 1, 2, 3 \cdots K. \tag{2.11}$$

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, the conditions (2.10) and (2.11) imply the following

$$\frac{\|\boldsymbol{\mathcal{T}}\|_\infty}{K} \le \mu_{\max} \sqrt{\kappa}, \tag{2.12}$$

where $\mu_{\max} := \max_i\{\mu_i\}$ and $\kappa := \max_i\{\kappa_i\}$. Here the condition (2.12) is a rough characterization of the mutual incoherence condition among different modes. Therefore, by comparing it with our condition (2.9), i.e.,

$$\frac{\|\boldsymbol{\mathcal{T}}\|_\infty}{K} \le \mu \sqrt{\kappa_k},$$

we can see that (2.9) strengthens the original matrix incoherence condition (2.8) and potentially implies a larger μ, especially when $\kappa \gg \kappa_k$. Obviously $\kappa \gg \kappa_k$ means that the tensor $\boldsymbol{\mathcal{X}}_0$ is somewhat "unbalanced" with respect to its ranks and incoherence for different modes. Hence, as suggested by our new mutual incoherence condition (2.9), the "balance" of the tensor $\boldsymbol{\mathcal{X}}_0$ does play an important role for exact recovery using the SNN model. Moreover, the importance of (2.9) has also been supported by our previous numerical results. From Figure 1, we have seen that for a well-balanced tensor $\boldsymbol{\mathcal{X}}_0$ with Tucker rank $(1, r, r)$, as r approaches 1, the performance of the SNN is significantly enhanced. With all other conditions being the same, the SNN model exactly recovers $\boldsymbol{\mathcal{X}}_0$ using much fewer observations than the Singleton model.

On the other hand, to demonstrate the effect of (2.9) on μ, we randomly generated a 3-way tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{100\times 100\times 100}$ 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 the core tensor $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{1\times r\times r}$ had entrees generated from an i.i.d Gaussian distribution, and each A_i was a random orthogonal matrix. In the first example shown in Figure 2, 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 2, we observe that the ratio $\frac{\|\boldsymbol{\mathcal{T}}\|_\infty/K}{\|\lambda_1 \boldsymbol{U}_1 \boldsymbol{V}_1^\top\|_\infty}$ grows gradually as r and κ are increased. This means that the gap between (2.8) and (2.9) becomes larger as $\boldsymbol{\mathcal{X}}_0$ get more unbalanced in terms of its Tucker ranks.

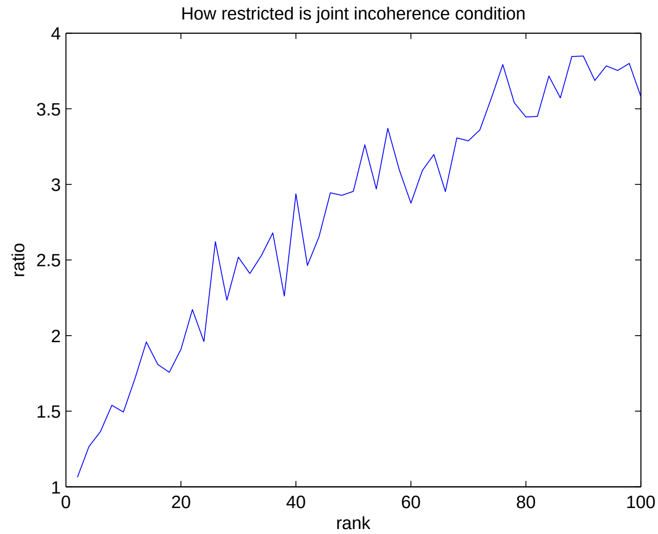

Figure 2: The average ratio $\frac{\|\sum_{i=1}^K \lambda_i \mathcal{A}_i^* \boldsymbol{U}_i \boldsymbol{V}_i^\top\|_\infty/K}{\|\lambda_1 \boldsymbol{U}_1 \boldsymbol{V}_1^\top\|_\infty}$ as a function of the rank r for randomly generated 3-way tensors $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{100\times 100\times 100}$ with Tucker rank $(1, r, r)$. For each rank $r \in [1, 100]$, we ran 10 independent trials and averaged their output ratios

It is also interesting to see how μ scales with the dimension n for $\boldsymbol{\mathcal{X}}_0 \in \mathbb{R}^{n\times n\times n}$ with different Tucker ranks. In Figure 3, we let n to increase from 5 to 250 and compare the cases of low Tucker ranks, i.e., $(2, 2, 2)$ and high Tucker ranks, i.e., $(2, n, n)$. We compute the smallest μ induced by (2.9), i.e.,

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

The left subplot in Figure 3 shows that μ scales well with n with a rate of $O(\log(n_k^{(1)}))$ for the low-rank tensor $\boldsymbol{\mathcal{X}}_0$. This is not surprising since each mode i is incoherent thus implies $\mu_i = O(\log(n_k^{(1)}))$. However, when the tensor $\boldsymbol{\mathcal{X}}_0$ is more unbalanced in terms of its Tucker ranks, μ scales badly as the dimensionality n grows. The right subplot indicates that μ grows slower than $\sqrt{n_k^{(1)}}$ in both cases.

As demonstrated in Figure 2 and 3, in the case where $\boldsymbol{\mathcal{X}}_0$ has unbalanced Tucker ranks, the condition (2.9) requires a larger μ, thus leading to a weaker ability for exact recovery. Therefore, (2.9) 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. Indeed, in Figure 1 we illustrate the difference between the SNN model and the Singleton model for recovering an incomplete tensor $\boldsymbol{\mathcal{X}}_0$ under different ranks and observation levels. As the rank r is increased, κ/κ_k gets larger and larger. Therefore, (2.9) becomes more restrictive and SNN minimization

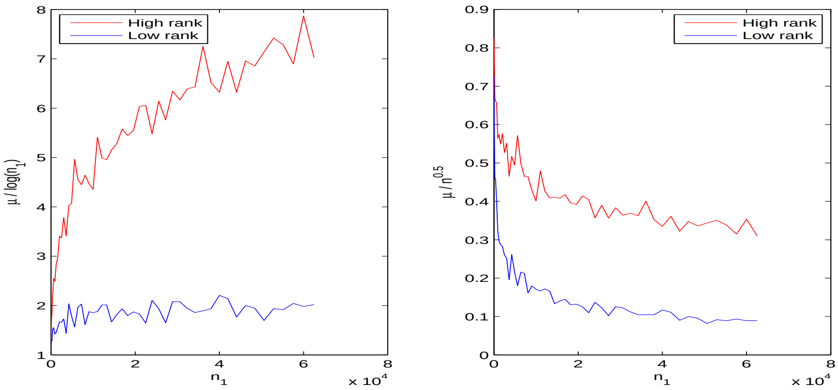

Figure 3: The smallest value of μ implied by (2.9) as a function of the dimensionality n for randomly generated 3-way tensors $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n\times n\times n}$ with Tucker rank $(2, r, r)$. We compare the low-rank case with $r = 2$, and the high-rank case with $r = n$. For each rank $n \in [5, 250]$, we ran 5 independent trials and averaged their output for $\|\boldsymbol{\mathcal{T}}\|_\infty$

is more likely to perform poorly. In particular, (2.9) suggests that for good recovery, the average overall incoherence should be on a par with the incoherence of the low-rank (kth) mode as measured by the infinity norm.

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, thus $\Omega = \Lambda \cup \Gamma$. Here we consider the following strongly convex version of this tensor model that is analogous to the matrix model (1.9).

$$\begin{aligned} \min_{\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{E}}} \quad & f(\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{E}}) := \sum_{i=1}^{K} \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* + \|\boldsymbol{\mathcal{E}}\|_1 + \frac{\tau}{2}\|\boldsymbol{\mathcal{X}}\|_F^2 + \frac{\tau}{2}\|\boldsymbol{\mathcal{E}}\|_F^2 \\ \text{s.t.} \quad & \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{E}}] = \boldsymbol{\mathcal{B}} \end{aligned} \tag{2.13}$$

Now we present our main theorem on conditions under which solving (2.13) yields the exact recovery of the true low-rank tensor $\boldsymbol{\mathcal{X}}_0$. The detailed proof is provided in the next section.

Theorem 2.2. *Suppose $\boldsymbol{\mathcal{X}}_0$ obeys the incoherence conditions (2.8)-(2.9) with parameter 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*

$$r_k \le C_r K^{-2} \frac{\rho\mu^{-1} n_k^{(2)}}{\log^2 n_k^{(1)}}, \quad \gamma \le C_\gamma \tag{2.14}$$

or equivalently

$$\rho \ge C_\rho r_k K^2 \frac{\log^2 n_k^{(1)}}{n_k^{(2)}}, \quad \gamma \le C_\gamma$$

and

$$\tau \le \frac{1}{2n_k^{(1)} n_k^{(2)} (1 + \frac{4}{\rho(1-2C_\gamma)}) \|\boldsymbol{\mathcal{B}}\|_F}, \tag{2.15}$$

solving (2.13) 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, C_ρ and C_γ are positive numbers.

As special cases of (2.13), the recovery guarantees for models (2.4) and (2.5) are naturally implied by Theorem 2.2 as in the following corollaries.

Corollary 2.3. *Suppose* $\boldsymbol{\mathcal{X}}_0$ *satisfies the incoherence conditions* (2.8)-(2.9) *with parameter* μ*, and* $\boldsymbol{\mathcal{E}}_0$ *has support uniformly distributed with probability* γ *and with arbitrary magnitude. Then provided that*

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

and

$$\lambda_i = \sqrt{n_i^{(1)}}$$

we can exactly recover $(\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{E}}_0)$ *via solving* (2.5) *with probability at least* $1 - Cn^{-3}$. *Here* C, C_r *and* C_γ *are all positive numbers.*

Corollary 2.3 holds from Theorem 2.2 with $\rho = 1$ and $\tau = 0$. It shows that the RPCA model (2.5) can exactly recover $(\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{E}}_0)$ with only $n_k^{(1)} n_k^{(2)}$ observations, while the order of the degrees of freedom in the underlying signal is $C n_k^{(1)} n_k^{(2)}$ and $C > 1$. The above bound is consistent with the best known result for matrix RPCA.

Corollary 2.4. *Suppose* $\boldsymbol{\mathcal{X}}_0$ *satisfies the incoherence conditions* (2.8)-(2.9) *with parameter* μ*, and observations are supported on* Ω*, which uniformly distributed with cardinality* m*. Then provided that* $m \geq \frac{K^2 \mu}{C_r} r_k n_k^{(1)} \log^2 n_k^{(1)}$ *and* $\lambda_i = \sqrt{n_i^{(1)}}$*, we can exactly recover* $\boldsymbol{\mathcal{X}}_0$ *via solving* (2.4) *with probability at least* $1 - Cn^{-3}$*. Here* C *and* C_r *are both positive numbers.*

Corollary 2.4 holds from Theorem 2.2 with $\gamma = 0$ and $\tau = 0$. Unlike Corollary 2.3 for the RPCA model, the above result is arguably suboptimal for the cases when $K \geq 4$. To illustrate, consider a K-way tensor with length n and rank r in each mode. Then the degree of freedom of this tensor is $O(r^k + rnk)$, which is substantially smaller than the number of measurements $O(rn^{K-1})$ required in Corollary 2.4. For $K \geq 4$, [27] proposes a square model, minimizing the nuclear norm of a more balanced embedding of the tensor, which only needs $O(r^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil})$ Gaussian measurements. However, when $K = 3$, to the best of our knowledge, at least $O(rn^2)$ measurements are required to recover a general low-rank tensor using any convex model that is computationally tractable.

3 Architecture of the Proof

3.1 Sampling Schemes and Model Randomness

Theorem 2.2 is established based on using a *Uniform sampling scheme without replacement* to choose a set of entrees Ω with cardinality m. However, in order to 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 ([5], [4]) 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 $\Omega \sim \text{Ber}(\rho)$ 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

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

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 2.2 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:

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

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 [4]), *if the recovery of* $(\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{E}}_0')$ *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 3.1. *With the tensor* $\boldsymbol{\mathcal{T}}$ *defined as in Definition 2.1, 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 [5] and (3.9) can be found in [4]. Note that all three inequalities are applied to the matricization on the kth mode where k is the low-rank mode.

Lemma 3.2. *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 (2.8)-(2.9), and Ω_k is the support Ω applied to the kth mode. 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 3.3. *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 (2.13) 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

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

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, \qquad (3.12)
\end{aligned}
$$

where the first inequality follows directly from 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

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

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 2.2:

Proof. We apply the *Golfing Scheme* similar to that in [23] to construct the dual certificate $\boldsymbol{Y}$ that satisfies

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

and verify the following condition for τ

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

.

[**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 $\Gamma_k \sim \mathrm{Ber}((1-2\gamma)\rho)$ as a union of sets of support Γ^j, i.e., $\Gamma_k = \bigcup_{j=1}^p \Gamma^j$ where $\Gamma^j \sim \mathrm{Ber}(q_j)$. Define $q_1 = q_2 = \frac{(1-2\gamma)\rho}{6}$ and $q_3 = \cdots = q_p = q$, which implies $q \ge C\rho/\log n_k^{(1)}$. 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 3.2 that

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

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

$$\|\boldsymbol{Z}^0\|_\infty \le \|\boldsymbol{\mathcal{T}}_{(k)}\|_\infty + \|\mathcal{P}_{T_k}[\boldsymbol{\mathcal{S}}_{(k)}]\|_\infty. \tag{3.17}$$

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

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

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 (2.9)

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

Therefore from (3.17) we have

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

$$\|\boldsymbol{Z}^0\|_F \leq \sqrt{n_k^{(1)} n_k^{(2)}}\|\boldsymbol{Z}^0\|_\infty \leq CK\sqrt{\mu r_k n_k^{(1)} \log n_k^{(1)}} \tag{3.19}$$

- 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 $\boldsymbol{\mathcal{S}}_{(k)} = \operatorname{sgn}(\boldsymbol{\mathcal{E}}_0)$ 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

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

when C_γ is sufficiently small.

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

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

provided C_ρ is sufficiently large.

- Last, we show that

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

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.,

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

[**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

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

We observe that

$$\|\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{E}}_0\|_F = \|(\mathcal{I} + \mathcal{P}_\Omega)[\boldsymbol{\mathcal{X}}_0] - \boldsymbol{\mathcal{B}}\|_F \le 2\|\boldsymbol{\mathcal{X}}_0\|_F + \|\boldsymbol{\mathcal{B}}\|_F. \tag{3.20}$$

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

$$\|\boldsymbol{\mathcal{X}}_0\|_F \le \frac{2}{(1-2\gamma)\rho} \left\|\mathcal{P}_{T_k} \mathcal{P}_{\Gamma_k} \mathcal{P}_{T_k} [(\boldsymbol{\mathcal{X}}_0)_{(k)}]\right\|_F = \frac{2}{(1-2\gamma)\rho} \left\|\mathcal{P}_{T_k} \mathcal{P}_{\Gamma_k} [\boldsymbol{\mathcal{B}}_{(k)}]\right\|_F \le \frac{2}{(1-2\gamma)\rho} \|\boldsymbol{\mathcal{B}}\|_F. \tag{3.21}$$

Therefore we have

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

and it suffices to have

$$\tau \leq \frac{1}{2n_k^{(1)} n_k^{(2)} \left(1 + \frac{4}{p_0(1-\gamma_s)}\right) \|\boldsymbol{\mathcal{B}}\|_F}, \tag{3.22}$$

□

4 Discussions

In this paper, we establish a theoretical bound for a RLRTC model, i.e., (2.13), based on SNN convexification. The model (2.13), 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. Our paper presents, to the best of our knowledge, the first rigorous study on theoretical guarantees for this model. Simulations show that, with λ's being fixed, the tensor model (2.13) performs much better when the target tensor $\boldsymbol{\mathcal{X}}_0$ has more balanced Tucker ranks. Thus we propose a new set of tensor incoherence conditions under which using high-order tensor models based on SNN minimization is plausible.

For the tensor completion problem, our results imply that, using SNN, $\Omega\left(rn^{K-1}\right)$ observed entries are sufficient to recover a K-way tensor with length n and rank r along each mode, which is recently proved to be tight by [29]. Also it is worth mentioning that the SNN-based models such as (2.13), despite being very popular, are not the only effective approach for solving the RLRTC problem. Recently, other models – e.g. [21, 17, 47, 2, 43], have been proposed with potentially better complexity bounds.

Furthermore, the bounds achieved in Theorem 2.2 do not explain why the SNN model (2.13) 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 model (2.13) and would be 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. *The Journal of Machine Learning Research 15*, 1 (2014), 2773–2832.

[2] Aswani, A. Positive low-rank tensor completion. *arXiv preprint arXiv:1412.0620* (2014).

[3] 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.

[4] Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM) 58*, 3 (2011), 11.

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

[6] 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.

[7] Cohen, S. B., and Collins, M. Tensor decomposition for fast latent-variable PCFG parsing. In *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 ℓ_1-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 *ISWC* (2009).

[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. *SIAM Journal on Matrix Analysis and Applications 35*, 1 (2014), 225–253.

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

[15] Harshman, R. A. Foundations of the parafac procedure: models and conditions for an explanatory multimodal factor analysis. *UCLA Working Papers in Phonetics 16* (1970).

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

[17] Jain, P., and Oh, S. Provable tensor factorization with missing data. In *NIPS* (2014).

[18] Jiang, B., Ma, S., and Zhang, S. Tensor principal component analysis via convex optimization. *Mathematical Programming 150*, 2 (2015), 423–457.

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

[20] Kreimer, N., S. A., and Sacchi, M. D. Nuclear norm minimization and tensor completion in exploration seismology. In *ICASSP* (2013).

[21] Kressner, D., Steinlechner, M., and Vandereycken, B. Low-rank tensor completion by riemannian optimization. *BIT Numerical Mathematics 54*, 2 (2014), 447–468.

[22] Li, N., and Li, B. Tensor completion for on-board compression of hyperspectral images. In *ICIP* (2010).

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

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

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

[26] 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.

[27] Mu, C., Huang, B., Wright, J., and Goldfarb, D. Square deal: Lower bounds and improved relaxations for tensor recovery. In *ICML* (2014).

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

[29] Oymak, S., Jalali, A., Fazel, M., Eldar, Y. C., and Hassibi, B. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory* (2015), to appear.

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

[31] Recht, B. A simpler approach to matrix completion. *The Journal of Machine Learning Research 12* (2011), 3413–3430.

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

[33] 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).

[34] 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).

[35] 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.

[36] 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.

[37] Sun, J.-T., Zeng, H.-J., Liu, H., Lu, Y., and Chen, Z. Cubesvd: a novel approach to personalized web search. In *WWW* (2005).

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

[39] Tomioka, R., Suzuki, T., Hayashi, K., and Kashima, H. Statistical performance of convex tensor decomposition. In *NIPS* (2011).

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

[41] Vasilescu, M. A. O., and Terzopoulos, D. Multilinear subspace analysis of image ensembles. In *CVPR* (2003).

[42] 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. *NIPS* (2009).

[43] Xu, Y., Hao, R., Yin, W., and Su, Z. Parallel matrix factorization for low-rank tensor completion. *Inverse Problems & Imaging 9*, 2 (2015), 601–624.

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

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

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

[47] Yuan, M., and Zhang, C.-H. On tensor completion via nuclear norm minimization. *arXiv preprint arXiv:1405.1773* (2014).

[48] Zhang, H., Cai, J.-F., Cheng, L., and Zhu, J. Strongly convex programming for exact matrix completion and robust principal component analysis. *Inverse Problems & Imaging 6*, 2 (2012).