# Tensor Robust Principal Component Analysis: Better recovery with atomic norm regularization

# Tensor Robust Principal Component Analysis: Better recovery with atomic norm regularization

Derek Driggs[*] Stephen Becker[†] Jordan Boyd-Graber[‡]

**Abstract**

This paper studies tensor-based Robust Principal Component Analysis (RPCA) using atomic-norm regularization. Given the superposition of a sparse and a low-rank tensor, we present conditions under which it is possible to exactly recover the sparse and low-rank components. Our results improve on existing performance guarantees for tensor-RPCA, including those for matrix RPCA. Our guarantees also show that atomic-norm regularization provides better recovery for tensor-structured data sets than other approaches based on matricization.

In addition to these performance guarantees, we study a nonconvex formulation of the tensor atomic-norm and identify a class of local minima of this nonconvex program that are globally optimal. We demonstrate the strong performance of our approach in numerical experiments, where we show that our nonconvex model reliably recovers tensors with ranks larger than all of their side lengths, significantly outperforming other algorithms that require matricization.

**Keywords:** Tensor completion, Matrix completion, Nonconvex optimization, Tensor rank, Principal component analysis

**AMS Subject Classification:** Primary 90C25; Secondary 15A69, 15A83.

# 1 Introduction

Tensors, or multidimensional arrays, are becoming increasingly prominent in data analysis and machine learning. Tensors were first used as tools for data analysis in the psychometics community, where researchers used tensor decompositions to study fMRI data sets that are more naturally represented as tensors than matrices [26, 44]. Since then, tensors have established a place in chemometrics, computer vision, compressed sensing, data mining, and higher-order statistics [37]. Tensors' ability to naturally represent the distributions of latent variable models has also made them important tools for learning a variety of latent variable models, including topic models, Gaussian mixture models, and neural networks, to name a

[*] Applied Mathematics and Theoretical Physics, University of Cambridge (d.driggs@cam.ac.uk)

[†] Applied Mathematics, University of Colorado Boulder (stephen.becker@colorado.edu)

[‡] Computer Science, iSchool, lsc, and umiacs, University of Maryland College Park (jbg@umiacs.umd.edu)

few [4, 25]. With modern data sets growing quickly in both size and complexity, tensor-based algorithms offer more natural approaches for analyzing multidimensional data.

Many tensor-based algorithms for data analysis are formulated as low-rank recovery problems. Low-rank tensor decompositions have been used for video processing [5, 39], topic modeling [3], blind source separation [14], and parameter estimation in latent variable models [4]. Low-rank matrix recovery problems have been studied extensively, and Robust Principal Component Analysis (RPCA) is one approach to this problem [12]. RPCA decomposes the superposition of a low-rank and a sparse matrix into their original components by solving a convex optimization program, thereby efficiently recovering a low-rank matrix from grossly corrupted measurements.

Recently, several authors have proposed tensor completion algorithms for low-rank recovery [8, 19, 39, 47, 42, 33], and only a few of these techniques have been extended to RPCA [5, 24, 36, 31]. Many of these algorithms do not work with tensors directly, but instead work with matrix representations of the higher-order data. Recent work has shown that representing tensors as matrices leads to sub-optimal performance [47], but there are few methods for low-rank tensor recovery that are tractable, have performance guarantees, and preserve the tensor's higher-order structure.

In this paper, we study the following model for tensor RPCA:

$$\min_{\mathcal{X},\mathcal{S}} \quad \|\mathcal{X}\|_* + \lambda\|\mathcal{S}\|_{\text{sum}} \quad \text{subject to:} \quad \mathcal{X} + \mathcal{S} = \mathcal{Z}$$

Our model uses the tensor atomic-norm: a higher-order generalization of the matrix nuclear-norm (see Section 2 for further details). We present the first performance guarantees for tensor RPCA using the tensor atomic norm. Our results improve on all existing recovery guarantees for tensor RPCA, corroborating recent literature suggesting that preserving the structure of multidimensional data sets allows for significantly improved recovery. The order-two case of our bounds also offer slightly improved recovery guarantees for matrix RPCA. We also discuss a nonconvex algorithm for solving our tensor RPCA program, and show that although our problem is nonconvex, all local minima are globally optimal. Our experiments show that our model can consistently recover tensors with full Tucker-rank (but not full CP-rank). This suggests that although our performance guarantees are sharp using the Tucker-rank as a metric, they are not yet sharp in terms of the CP-rank.

The rest of this paper is outlined as follows. In Section 2, we provide an overview of some analytic and algebraic properties of tensors. In Section 3, we present recovery guarantees for our formulation of tensor RPCA using atomic-norm regularization. We also compare these guarantees to existing results for tensor RPCA. Section 4 contains the proof of these recovery guarantees. While the tensor atomic norm is convex, it is in general computationally intractable, so in Section 5, we discuss a nonconvex representation of the tensor atomic norm that can be seen as a higher-order generalization of the Burer-Monteiro factorization approach that is popular in low-rank matrix recovery algorithms. This formulation has been used previously for tensor completion, and we show how it can be extended to tractably find stationary points of the tensor RPCA program. We also show that all local minima of our nonconvex program are globally optimal. We discuss how tensor RPCA cam be used for training latent variable models in Section 6, using topic modeling as a motivating example. Finally, we present numerical experiments in Section 7 that demonstrate the efficacy of our approach.

## 2 Tensor Preliminaries

This section introduces definitions and properties of tensors. A more complete review of this information can be found in [17], [26], and [47]. We focus on order-three tensors to simplify notation, but our results can easily be extended to arbitrary orders.

Let $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ be an order-three tensor with side lengths $d_1, d_2, d_3$. The *fibers* of $\mathcal{X}$ *along its $k^{th}$ mode* are the vectors obtained by holding all but one of the indices of $\mathcal{X}$ fixed and varying the $k^{th}$ index. In some cases, it is useful to *matricize* a tensor, so that it is represented as a matrix. The matricization of an order-three tensor $\mathcal{X}$ along its $k^{th}$ mode is denoted $\mathcal{X}_{(k)}$ and is formed by taking the mode-$k$ fibers of $\mathcal{X}$ and making them the columns of $\mathcal{X}_{(k)}$.

The tensor product, which we denote $\otimes$, is defined so that if $\mathcal{X} = u \otimes v \otimes w$, then $\mathcal{X}_{i_1,i_2,i_3} = u_{i_1} v_{i_2} w_{i_3}$. The tensor product generalizes the outer product, so $u \otimes v = uv^T$. Every tensor $\mathcal{X}$ admits a *CP-decomposition* (CPD) of the form

$$\mathcal{X} = \sum_{r=1}^{R} (a_r \otimes b_r \otimes c_r) = \sum_{r=1}^{R} \gamma_r (u_r \otimes v_r \otimes w_r), \tag{1}$$

where the vectors $u_r, v_r$, and $w_r$ have unit-norm and $\gamma_r = \|a_r\|\|b_r\|\|c_r\|$. When $R$ is minimal, we call $R$ the *CP-rank* of $\mathcal{X}$. The matrices $A, B, C$ that have $a_r, b_r, c_r$ as their columns, respectively, are *factor matrices* of $\mathcal{X}$. It is also sometimes convenient to use Kruskal's notation $\mathcal{X} = [\![A, B, C]\!] = [\![\gamma; U, V, W]\!]$ to denote the decomposition in (1).

Matrices can act on a tensor through multiplication. For a tensor $\mathcal{X} = [\![A, B, C]\!] \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, multiplication by the matrices $M_i \in \mathbb{R}^{k_i \times d_i}$, $i = 1, 2, 3$ is defined as follows:

$$(M_1, M_2, M_3) \cdot \mathcal{X} = [\![M_1 A, M_2 B, M_3 C]\!] \in \mathbb{R}^{k_1 \times k_2 \times k_3},$$

and multiplication between the factor matrices is canonical matrix multiplication. We choose to use the notation of [38], but this is multiplication is the same as $\mathcal{X} \times_1 M_1 \times_2 M_2 \times_3 M_3$ using the notation of [26].

We also use the *Khatri-Rao* product, which we denote $\odot$. For two matrices $A \in \mathbb{R}^{m \times n}$. $B \in \mathbb{R}^{p \times n}$ with the same number of columns, we have

$$A \odot B \overset{\text{def}}{=} \begin{pmatrix} a_{1,1}\mathbf{b_1} & \cdots & a_{1,n}\mathbf{b_n} \\ \vdots & \ddots & \vdots \\ a_{m,1}\mathbf{b_1} & \cdots & a_{m,n}\mathbf{b_n} \end{pmatrix} \in \mathbb{R}^{m \cdot p \times n},$$

where $\mathbf{b_i}$ is the $i^{th}$ column of $B$. A matricized tensor can be expressed neatly using the Khatri-Rao product [26].

The *Tucker-rank* of $\mathcal{X}$ is the tuple $\big(\text{rank}(\mathcal{X}_{(1)}), \text{rank}(\mathcal{X}_{(2)}), \text{rank}(\mathcal{X}_{(3)})\big)$. We can bound the CP-rank of a third-order tensor using a weighted average of the components of its Tucker-rank:

$$\overline{r}(\mathcal{X}) \overset{\text{def}}{=} \sqrt{\frac{r_1 r_2 d_3 + r_1 r_3 d_2 + r_2 r_3 d_1}{d_1 + d_2 + d_3}}, \quad \text{where } r_i \overset{\text{def}}{=} \text{rank}(\mathcal{X}_{(i)}).$$

If $d_1 = d_2 = d_3 = d$ then $\overline{r} \leq d$. It has been established that the CP-rank $R \in [\overline{r}, \overline{r}^2]$ [47].

## 2.1 Coherence

Exact recovery of a tensor through our RPCA model relies on the tensor having low coherence. We adopt the measures of tensor coherence introduced in [47]. Recall that the coherence of an $r$-dimensional linear subspace span$(U)$ of $\mathbb{R}^k$ is defined to be [13, 47]

$$\mu(U) \overset{\text{def}}{=} \frac{k}{r} \max_{1\le i\le k} \|\mathcal{P}_U e_i\|^2 = \frac{\max_{1\le i\le k} \|\mathcal{P}_U e_i\|^2}{k^{-1}\sum_{i=1}^{k} \|\mathcal{P}_U e_i\|^2},$$

where $\mathcal{P}_U$ is the projection onto span$(U)$. For a tensor $\mathcal{X} = [\![A, B, C]\!] \in \mathbb{R}^{d_1\times d_2\times d_3}$, we define one measure of coherence as

$$\mu(\mathcal{X}) \overset{\text{def}}{=} \max\{\mu(A), \mu(B), \mu(C)\}.$$

We can also interpret $\mu(\mathcal{X})$ as the maximum coherence of the column spaces of each matricization of $\mathcal{X}$. Another measure of coherence, for $\mathcal{X} = [\![A, B, C]\!]$, is

$$\alpha(\mathcal{X}) \overset{\text{def}}{=} \sqrt{d_1 d_2 d_3/\overline{r}(\mathcal{X})}\|\mathcal{W}\|_{\max}, \tag{2}$$

where $\mathcal{W} = [\![W_1, W_2, W_3]\!]$ satisfies $\mathcal{W} = [\![\mathcal{P}_A W_1, \mathcal{P}_B W_2, \mathcal{P}_C W_3]\!]$, $\|\mathcal{W}\| = 1$, $\langle \mathcal{X}, \mathcal{W}\rangle = \|\mathcal{X}\|_*$, and $\|\mathcal{W}\|_{\max}$ is the largest entry of $\mathcal{W}$ in absolute value (cf. (4)).

We assume that the low-rank component $\mathcal{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$ has low coherence, so it satisfies $\mu(\mathcal{X}) \le \mu_0$ and $\alpha(\mathcal{X}) \le \alpha_0$. These coherence bounds are not especially more restrictive than the coherence bounds required for low-rank matrix recovery. A bounded coherence ensures that the low-rank component of $\mathcal{X}$ is not sparse, so it can be separated from the sparse component. $\mu_0$ bounds the coherence of each matricization of $\mathcal{X}$, and $\alpha_0$ provides a uniform bound on these coherences similar to the uniform bound used in [24].

## 2.2 Projection Operators

In the proof of our main result, we use projection operators that act on order-three tensors. Let $U, V, W$ be matrices. With $\mathcal{X} = [\![A, B, C]\!]$, define the projection operator $\mathcal{P}_{U,V,W} : \mathcal{X} \mapsto [\![\mathcal{P}_U(A), \mathcal{P}_V(B), \mathcal{P}_W(C)]\!]$, where $\mathcal{P}_U$, for example, projects matrices onto the column space of $U$, and $\mathcal{P}_{U^\perp}$ projects onto the orthogonal complement of the column space of $U$. For convenience, we adopt the notation of [47] to define the following projections:

$$\begin{aligned}
\mathcal{P}^0_{\mathcal{X}} &\overset{\text{def}}{=} \mathcal{P}_{U,V,W}\\
\mathcal{P}_{\mathcal{X}} &\overset{\text{def}}{=} \mathcal{P}_{U,V,W} + \mathcal{P}_{U^\perp,V,W} + \mathcal{P}_{U,V^\perp,W} + \mathcal{P}_{U,V,W^\perp}\\
\mathcal{P}_{\mathcal{X}_1} &\overset{\text{def}}{=} \mathcal{P}_{U^\perp,V^\perp,W}\\
\mathcal{P}_{\mathcal{X}_2} &\overset{\text{def}}{=} \mathcal{P}_{U^\perp,V,W^\perp}\\
\mathcal{P}_{\mathcal{X}_3} &\overset{\text{def}}{=} \mathcal{P}_{U,V^\perp,W^\perp}\\
\mathcal{P}_{\mathcal{X}_4} &\overset{\text{def}}{=} \mathcal{P}_{U^\perp,V^\perp,W^\perp}\\
\mathcal{P}_{\mathcal{X}^\perp} &\overset{\text{def}}{=} \mathcal{P}_{\mathcal{X}_1} + \mathcal{P}_{\mathcal{X}_2} + \mathcal{P}_{\mathcal{X}_3} + \mathcal{P}_{\mathcal{X}_4} = I - \mathcal{P}_{\mathcal{X}}.
\end{aligned}$$

These are orthogonal projections, i.e., if $\mathcal{P}$ is any of the above, then $\mathcal{P}^2 = \mathcal{P}$ and $\mathcal{P} = \mathcal{P}^\top$.

Let $\mathcal{S}$ be a sparse tensor. The *support* of $\mathcal{S}$, which we denote $\Omega$, is the set of indices corresponding to the nonzero entries of $\mathcal{S}$. We use $\mathcal{P}_\Omega$ to denote the projection onto the support of $\mathcal{S}$, i.e.,

$$(\mathcal{P}_\Omega(\mathcal{X}))_{i_1,i_2,i_3} = \begin{cases} \mathcal{X}_{i_1,i_2,i_3} & (i_1,i_2,i_3) \in \Omega, \\ 0 & (i_1,i_2,i_3) \notin \Omega. \end{cases}$$

## 2.3 Norms for Tensors and Operators on Tensors

In the proof of our main result, we analyze linear operators acting on order-three tensors and their operator norms, which we define with respect to the usual Euclidean inner product. Let

$$\langle \mathcal{X}, \mathcal{Y} \rangle \overset{\text{def}}{=} \sum_{(i,j,k)\in[d_1]\times[d_2]\times[d_3]} \mathcal{X}_{i,j,k}\mathcal{Y}_{i,j,k}.$$

The Frobenius (or Hilbert-Schmidt) norm is the induced norm $\|\mathcal{X}\|_F^2 = \langle \mathcal{X}, \mathcal{X} \rangle$. Let $\mathcal{Q} : \mathbb{R}^{d_1\times d_2\times d_3} \to \mathbb{R}^{d_1'\times d_2'\times d_3'}$ be a linear operator. We define the norm of $\mathcal{Q}$ as the operator norm

$$|||\mathcal{Q}||| \overset{\text{def}}{=} \sup_{\substack{\mathcal{X}\in\mathbb{R}^{d_1\times d_2\times d_3} \\ \|\mathcal{X}\|_F\le 1}} \|\mathcal{Q}(\mathcal{X})\|_F. \tag{3}$$

Finally, we need higher-dimensional generalizations of the matrix $\ell_1$ and $\ell_\infty$ norms. Write $[d_1]$ as shorthand for the list $(1, 2, \ldots, d_1)$. Then for a tensor $\mathcal{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$, we can define

$$\|\mathcal{X}\|_{\text{sum}} \overset{\text{def}}{=} \sum_{(i,j,k)\in[d_1]\times[d_2]\times[d_3]} |\mathcal{X}_{i,j,k}|,$$

and

$$\|\mathcal{X}\|_{\max} \overset{\text{def}}{=} \max_{(i,j,k)\in[d_1]\times[d_2]\times[d_3]} |\mathcal{X}_{i,j,k}|.$$

The *subdifferential* of a function $f$ at a point $x$ is the set $\partial f(x) \overset{\text{def}}{=} \{d \mid f(y) \ge f(x) + \langle d, y - x\rangle \; \forall y \in \text{dom}(f)\}$. The subdifferential of $\mathcal{X} \mapsto \|\mathcal{X}\|_{\text{sum}}$ reduces to

$$\partial\|\mathcal{X}\|_{\text{sum}} = \{\text{sgn}(\mathcal{X}) + \mathcal{F} \mid \mathcal{F} = \mathcal{P}_{\Omega^\perp}\mathcal{F},\ \|\mathcal{F}\|_{\max} \le 1\} \quad (\Omega = \text{supp}(\mathcal{X}))$$

where $\text{sgn}(\mathcal{X})$ computes the sign of $\mathcal{X}$ element-wise, with $\text{sgn}(0) = 0$. Note $\mathcal{F} = \mathcal{P}_{\Omega^\perp}\mathcal{F} \iff \mathcal{P}_\Omega\mathcal{F} = 0$.

The tensor *atomic norm* (also known as the tensor *nuclear norm*), which we denote as $\|\cdot\|_*$, is defined as follows:

$$\|\mathcal{X}\|_* = \min\left\{ \sum_{r=1}^R |\gamma_r| : \mathcal{X} = \sum_{r=1}^R \gamma_r(u_r \otimes v_r \otimes w_r),\ \|u_r\| = \|v_r\| = \|w_r\| = 1 \right\}$$

In the matrix case, the atomic norm is the matrix nuclear norm, which is equal to the sum of the singular values of a matrix. Unlike for matrices, the decomposition that realizes the minimum above is not necessarily the minimal-rank CPD of $\mathcal{X}$. We call a decomposition that does achieve the minimum an *atomic decomposition* of $\mathcal{X}$, and such a decomposition always exists [17]. The corresponding number of terms $R$ in an atomic decomposition is *atomic rank*, as we use in Section 5.1. The tensor atomic norm can be interpreted as the $\ell_1$-norm of the weights in its atomic decomposition. Roughly speaking, atomic-norm regularization encourages these weights to tend toward zero, promoting low-rank solutions just as $\ell_1$-norm regularization promotes sparsity.

The dual to the atomic norm is the spectral norm, $\|\cdot\|$, which is defined as [17, 47]:

$$\|\mathcal{X}\| = \max_{\|u\|=\|v\|=\|w\|=1} \langle \mathcal{X}, u \otimes v \otimes w \rangle$$

There always exists a unit-Euclidean-norm tensor that maximizes this inner-product, so the maximum is well-defined [17]. Furthermore, there always exists a *dual tensor* [47] where

$$\mathcal{W} \text{ is dual to } \mathcal{X} \iff \big(\|\mathcal{W}\| = 1,\ \mathcal{P}^0_{\mathcal{X}}\mathcal{W} = \mathcal{W} \text{ and } \langle \mathcal{X}, \mathcal{W}\rangle = \|\mathcal{X}\|_*\big) \tag{4}$$

The spectral and atomic norms satisfy $\|\mathcal{X}\| \le \|\mathcal{X}\|_F \le \|\mathcal{X}\|_*$. To see this, let $\mathcal{X} = \sum_{r=1}^{R} \gamma_r (u_r \otimes v_r \otimes w_r)$ be an atomic decomposition of $\mathcal{X}$. By the triangle inequality,

$$\|\mathcal{X}\|_F \le \sum_{r=1}^{R} \gamma_r \|u_r \otimes v_r \otimes w_r\|_F = \sum_{r=1}^{R} \gamma_r = \|\mathcal{X}\|_*. \tag{5}$$

The inequality $\|\mathcal{X}\| \le \|\mathcal{X}\|_F$ follows from the fact that the spectral norm is dual to the atomic norm.

The proof of our main result relies on the following partial characterization of the subdifferential of the tensor atomic norm, which can be found in [47]:

$$\partial\|\mathcal{X}\|_* \supset \{\mathcal{W} + \mathcal{P}_{\mathcal{X}^\perp}\mathcal{W}^\perp : \|\mathcal{W}^\perp\| \le \tfrac{1}{2},\ \mathcal{W} \text{ is dual to } \mathcal{X}\}.$$

# 3 Main Result

Given the sum of a low-rank tensor $\mathcal{X}$ and a sparse tensor $\mathcal{S}$, $\mathcal{Z} = \mathcal{X} + \mathcal{S}$, we would like to recover $\mathcal{X}$ and $\mathcal{S}$ by solving the following program:

$$\min_{\mathcal{X}', \mathcal{S}'} \quad \|\mathcal{X}'\|_* + \lambda \|\mathcal{S}'\|_{\text{sum}} \quad \text{subject to:} \quad \mathcal{X}' + \mathcal{S}' = \mathcal{Z} \tag{6}$$

Theorem 1 provides conditions under which (6) recovers $\mathcal{X}$ and $\mathcal{S}$ exactly with high probability.

**Theorem 1.** *Suppose tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ satisfies $\mu(\mathcal{X}) \le \mu_0$ and $\alpha(\mathcal{X}) \le \alpha_0$. Let $\mathcal{S} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ have a support set $\Omega$ that is uniformly distributed among all sets of cardinality $m$, and let $n = d_1 d_2 d_3 - m$. There then exists a positive constant $c$ so that (6) with $\lambda = (d_1 + d_2 + d_3)^{-1/2}$ exactly recovers $\mathcal{X}$ and $\mathcal{S}$ with probability $1 - (d_1 + d_2 + d_3)^{-1-c}$, provided that*

$$\overline{r}(\mathcal{X}) \le \rho_r \left( \frac{n}{(d_1 + d_2 + d_3)\log(n)\alpha_0^4 \mu_0^2} \right)^{1/2} \quad \text{and} \quad m \le \rho_s d_1 d_2 d_3. \tag{7}$$

*where $\rho_r, \rho_s$ are numerical constants.*

Our guarantees improve on existing bounds for tensor RPCA, and they are also an improvement over performing matrix RPCA on the matricized tensor [12].

Furthermore, these bounds are near optimal in the following sense. For simplicity, consider these bounds in the regime $d_1 = d_2 = d_3 = d$, $r_1 = r_2 = r_3 = r$ and $m \ll d^3$. The rank-bound in equation (7) is then

$$r \leq \rho_r \left( \frac{d}{\log(d)^{-1/2} \alpha_0^2 \mu_0} \right). \tag{8}$$

The maximum possible Tucker-rank is $r_1 = r_2 = r_3 = d$, so our program allows for exact recovery of a tensor with full Tucker-rank to within a factor of $\log(d)^{-3/2}$. In fact, our numerical experiments in Section 7 indicate that our nonconvex reformulation of (6) can often recover tensors of full Tucker-rank (but low CP-rank). These guarantees, along with our numerical experiments, suggest that the Tucker-rank is often a poor measure of the complexity of tensors when compared to the CP-rank.

Outside of this work, the best guarantees (in many regimes) for a tensor RPCA are from [24]. The relevant theorem from this paper is listed below for the case $K = 3$ and $d_1 = d_2 = d_3 = d$:

**Theorem 2** (Thm. 1 in [24])**.** *Consider the following sum-of-nuclear-norms model for tensor RPCA:*

$$\min_{\mathcal{X}', \mathcal{S}'} \quad d \sum_{i=1}^{3} \|\mathcal{X}'_{(i)}\|_* + \|\mathcal{S}'\|_1, \quad \text{subject to:} \quad \mathcal{X}' + \mathcal{S}' = \mathcal{Z}, \tag{9}$$

*Let $\mathcal{X}$ have Tucker-rank $(r_1, r_2, r_3)$ and let $S$ have support set $\Omega$ that is uniformly distributed among all sets of cardinality $m$. Then there exists a constant $C$ such that* (9) *exactly recovers $\mathcal{X}$ and $\mathcal{S}$ with probability $1 - Cd^{-3}$ as long as*

$$r_k \leq C_r \frac{\iota_0^{-1} d}{\log^2 d}, \quad \textit{and} \quad m \leq \rho d^3, \quad (\forall\, k = 1, 2, 3), \tag{10}$$

*for some constants $C_r$ and $\rho$, and incoherence parameter $\iota_0$.*

Comparing (8) with (10), we see that our bound is less restrictive on the rank of the tensor by a logarithmic factor. The incoherence parameter $\iota_0$ used in [24] satisfies $\mu_0 \leq \iota_0$, so this improvement is valid no matter the coherence of $\mathcal{X}$. More importantly, Theorem 2 bounds the maximum component of the Tucker-rank, while our result bounds a weighted average of all the components, so it is less restrictive in this sense as well.

While the guarantees of Theorem 2 are the best existing bounds in many regimes, there are other results that are stronger in special cases. [5] develop a nonconvex approach to tensor RPCA with associated convergence and performance guarantees. Their results are stronger than those in Theorem 2 when $\mathcal{S}$ is block-sparse. Because our model assumes that the support of $\mathcal{S}$ is uniformly distributed, direct comparisons with [5] are difficult. [31] offer another set of guarantees for tensor RPCA based on the tubal rank of a tensor (see also [32]). Because the results in [31] set bounds on the tubal rank of the low-rank component, it is difficult to directly compare these bounds to others, but these results are similar to those in Theorem 2. There are also some works that present non-convex algorithms for decomposing a tensor into low-rank and sparse components in the presence of noise [11, 41]. While the recovery guarantees of these papers are similar to ours, they are not directly applicable to the noiseless regime.

# 4 Proof of Theorem 1

Our proof of Theorem 1 is similar to the proof for the matrix case [12]. We use a partial characterization of the subdifferential of the tensor atomic norm to formulate a dual certificate that ensures exact recovery. We then prove that such a certificate exists with high probability provided that the rank of the low-rank component and the sparsity of the sparse component satisfy the bounds given in Theorem 1.

Lemma 3 establishes our dual certificate. It relies on the condition that $|||\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}||| < \frac{1}{2}$; we discuss conditions under which $|||\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}|||$ is bounded with high probability in Section 4.3.

**Lemma 3.** *Suppose $|||\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}||| < \frac{1}{2}$ and $\lambda \in (0,1)$. Then $(\mathcal{X}, \mathcal{S})$ is the unique solution to* (6) *if $\mathcal{Z} = \mathcal{X} + \mathcal{S}$ and there exists tensors $\mathcal{W}^\perp, \mathcal{D}$ and $\mathcal{F}$ satisfying*

$$\mathcal{W} + \mathcal{P}_{\mathcal{X}^\perp} \mathcal{W}^\perp = \lambda(\operatorname{sgn}(\mathcal{S}) + \mathcal{F} + \mathcal{P}_\Omega \mathcal{D}) \tag{11}$$

*where $\|\mathcal{W}^\perp\| \leq \frac{1}{4}$, $\mathcal{P}_\Omega \mathcal{F} = 0$, $\|\mathcal{F}\|_{\max} \leq \frac{1}{4}$, $\|\mathcal{P}_\Omega \mathcal{D}\|_F \leq \frac{1}{8}$, and $\Omega = \operatorname{supp}(\mathcal{S})$. (As before and throughout this paper, $\mathcal{W}$ satisfies $\mathcal{P}^0_\mathcal{X} \mathcal{W} = \mathcal{W}$, $\|\mathcal{W}\| = 1$, and $\langle \mathcal{W}, \mathcal{X} \rangle = \|\mathcal{X}\|_*$, i.e. $\mathcal{W}$ is dual to $\mathcal{X}$.)*

*Proof.* Let $\Delta$ be a perturbation away from the supposed optimal point $\mathcal{X}$, so $(\mathcal{X} + \Delta, \mathcal{S} - \Delta)$ is a feasible point of (6). Let $\mathcal{W}$ be dual to $\mathcal{X}$ and $\|\mathcal{W}^\perp\| \leq \frac{1}{2}$, so $\mathcal{W} + \mathcal{P}_{\mathcal{X}^\perp} \mathcal{W}^\perp \in \partial \|\mathcal{X}\|_*$, and let $\operatorname{sgn}(\mathcal{S}) + \mathcal{F} \in \partial \|\mathcal{S}\|_{\text{sum}}$.

We then have

$$\begin{aligned}\|\mathcal{X} + \Delta\|_* + \lambda \|\mathcal{S} - \Delta\|_{\text{sum}} \geq \|\mathcal{X}\|_* + &\langle \Delta, \mathcal{W} + \mathcal{P}_{\mathcal{X}^\perp} \mathcal{W}^\perp \rangle \\ &+ \lambda \left( \|\mathcal{S}\|_{\text{sum}} - \langle \Delta, \operatorname{sgn}(\mathcal{S}) + \mathcal{F} \rangle \right)\end{aligned}$$

By the duality of the tensor atomic and spectral norms, there exists a $\mathcal{W}^\perp$ satisfying $\|\mathcal{W}^\perp\| = \frac{1}{2}$ and $\langle \mathcal{P}_{\mathcal{X}^\perp} \Delta, \mathcal{W}^\perp \rangle = \frac{1}{2} \|\mathcal{P}_{\mathcal{X}^\perp} \Delta\|_*$. Similarly, we can choose $\mathcal{F}$ so that $\langle \mathcal{F}, \Delta \rangle = -\frac{1}{2} \|\mathcal{P}_{\Omega^\perp} \Delta\|_{\text{sum}}$. Hence,

$$\begin{aligned}&\|\mathcal{X} + \Delta\|_* - \|\mathcal{X}\|_* + \lambda(\|\mathcal{S} - \Delta\|_{\text{sum}} - \|\mathcal{S}\|_{\text{sum}}) \\ &\quad \geq \frac{1}{2} (\|\mathcal{P}_{\mathcal{X}^\perp} \Delta\|_* + \lambda \|\mathcal{P}_{\Omega^\perp} \Delta\|_{\text{sum}}) + \langle \Delta, \mathcal{W} - \lambda \operatorname{sgn}(\mathcal{S}) \rangle\end{aligned}$$

We would like to show that the right side of this inequality is positive unless $\Delta = 0$. To this end, we bound the magnitude of the last term. Using equation (11),

$$\begin{aligned}\left| \langle \Delta, \mathcal{W} - \lambda \operatorname{sgn}(\mathcal{S}) \rangle \right| &= |\langle -\mathcal{P}_{\mathcal{X}^\perp} \mathcal{W}^\perp + \lambda \mathcal{F} + \lambda \mathcal{P}_\Omega \mathcal{D}, \Delta \rangle| \\ &\leq |\langle \mathcal{P}_{\mathcal{X}^\perp} \mathcal{W}^\perp, \Delta \rangle| + \lambda |\langle \mathcal{F}, \Delta \rangle| + \lambda |\langle \mathcal{P}_\Omega \mathcal{D}, \Delta \rangle| \\ &\leq \|\mathcal{W}^\perp\| \|\mathcal{P}_{\mathcal{X}^\perp} \Delta\|_* + \lambda \|\mathcal{F}\|_{\max} \|\mathcal{P}_{\Omega^\perp} \Delta\|_{\text{sum}} \\ &\quad + \lambda \|\mathcal{P}_\Omega \mathcal{D}\|_F \|\mathcal{P}_\Omega \Delta\|_F\end{aligned}$$

$$< \frac{1}{4}\left(\|\mathcal{P}_{\mathcal{X}^\perp}\Delta\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}\Delta\|_{\text{sum}}\right) + \frac{\lambda}{8}\|\mathcal{P}_\Omega\Delta\|_F,$$

where we used the facts that $\mathcal{P}_{\mathcal{X}^\perp}$ and $\mathcal{P}_\Omega$ are self-adjoint and $\mathcal{F}$ is supported on $\Omega^\perp$. This yields

$$\begin{aligned}&\|\mathcal{X}+\Delta\|_* - \|\mathcal{X}\|_* + \lambda(\|\mathcal{S}-\Delta\|_{\text{sum}} - \|\mathcal{S}\|_{\text{sum}})\\ &\quad > \frac{1}{4}\left(\|\mathcal{P}_{\mathcal{X}^\perp}\Delta\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}\Delta\|_{\text{sum}}\right) - \frac{\lambda}{8}\|\mathcal{P}_\Omega\Delta\|_F\end{aligned}$$

The last term can be bounded.

$$\begin{aligned}\|\mathcal{P}_\Omega\Delta\|_F &= \|\mathcal{P}_\Omega(\mathcal{P}_{\mathcal{X}} + \mathcal{P}_{\mathcal{X}^\perp})\Delta\|_F\\ &\le \|\mathcal{P}_\Omega\mathcal{P}_{\mathcal{X}}\Delta\|_F + \|\mathcal{P}_\Omega\mathcal{P}_{\mathcal{X}^\perp}\Delta\|_F\\ &\le \frac{1}{2}\|\Delta\|_F + \|\mathcal{P}_{\mathcal{X}^\perp}\Delta\|_F\\ &\le \frac{1}{2}\|\mathcal{P}_\Omega\Delta\|_F + \frac{1}{2}\|\mathcal{P}_{\Omega^\perp}\Delta\|_F + \|\mathcal{P}_{\mathcal{X}^\perp}\Delta\|_F,\end{aligned}$$

where we have used the facts that $|||\mathcal{P}_\Omega\mathcal{P}_{\mathcal{X}}||| \le \frac{1}{2}$, $|||\mathcal{P}_\Omega||| \le 1$, and $|||\mathcal{P}_{\mathcal{X}}||| \le 1$. Hence,

$$\|\mathcal{P}_\Omega\Delta\|_F \le \|\mathcal{P}_{\Omega^\perp}\Delta\|_F + 2\|\mathcal{P}_{\mathcal{X}^\perp}\Delta\|_F.$$

We now have

$$\begin{aligned}\|\mathcal{X}+\Delta\|_* - \|\mathcal{X}\|_* +&\lambda(\|\mathcal{S}-\Delta\|_{\text{sum}} - \|\mathcal{S}\|_{\text{sum}})\\ &> \frac{1}{4}\left(\|\mathcal{P}_{\mathcal{X}^\perp}\Delta\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}\Delta\|_{\text{sum}}\right) - \frac{\lambda}{8}\big(\|\mathcal{P}_{\Omega^\perp}\Delta\|_F\\ &\qquad + 2\|\mathcal{P}_{\mathcal{X}^\perp}\Delta\|_F\big)\\ &\ge \frac{1}{4}\left(1-\lambda\right)\|\mathcal{P}_{\mathcal{X}^\perp}\Delta\|_* + \frac{\lambda}{8}\|\mathcal{P}_{\Omega^\perp}\Delta\|_{\text{sum}}.\end{aligned}$$

The final inequality follows from the fact that for any tensor $\mathcal{T}$, $\|\mathcal{T}\|_{\text{sum}} \ge \|\mathcal{T}\|_F$ and $\|\mathcal{T}\|_* \ge \|\mathcal{T}\|_F$ (see (5)). This shows that the perturbation $\Delta$ leads to a strict increase in the objective, unless $\Delta = 0$. □

In summary, to ensure exact recovery, it suffices to find a tensor $\mathcal{W}^\perp$ satisfying

$$\begin{cases}\mathcal{P}_{\mathcal{X}^\perp}\mathcal{W}^\perp = \mathcal{W}^\perp,\\ \|\mathcal{W}^\perp\| < \frac{1}{4},\\ \|\mathcal{P}_\Omega(\mathcal{W} - \lambda\text{sgn}(\mathcal{S}) + \mathcal{W}^\perp)\|_F \le \frac{\lambda}{8},\\ \|\mathcal{P}_{\Omega^\perp}(\mathcal{W} + \mathcal{W}^\perp)\|_{\max} < \frac{\lambda}{4},\end{cases} \tag{12}$$

where we have used (11) to state our conditions on $\mathcal{F}$ in terms of $\mathcal{W}^\perp$.

Instead of proving the existence of $\mathcal{W}^\perp$ directly, we find tensors $\mathcal{W}^{\mathcal{L}}$ and $\mathcal{W}^{\mathcal{S}}$ satisfying $\mathcal{P}_{\mathcal{X}^\perp}(\mathcal{W}^{\mathcal{L}} + \mathcal{W}^{\mathcal{S}}) = \mathcal{W}^{\mathcal{L}} + \mathcal{W}^{\mathcal{S}}$, $\mathcal{P}_\Omega\mathcal{W}^{\mathcal{S}} = \lambda\text{sgn}(\mathcal{S})$, and the following:

$$\begin{cases}(a)\quad \|\mathcal{W}^{\mathcal{L}}\| < \frac{1}{8},\\ (b)\quad \|\mathcal{P}_\Omega(\mathcal{W} + \mathcal{W}^{\mathcal{L}})\|_F \le \frac{\lambda}{8},\\ (c)\quad \|\mathcal{P}_{\Omega^\perp}(\mathcal{W} + \mathcal{W}^{\mathcal{L}})\|_{\max} < \frac{\lambda}{8},\end{cases} \tag{13}$$

$$\begin{cases} (d) \quad \|\mathcal{W}^{\mathcal{S}}\| < \frac{1}{8}, \\ (e) \quad \|\mathcal{P}_{\Omega^\perp}\mathcal{W}^{\mathcal{S}}\|_{\max} < \frac{\lambda}{8}. \end{cases} \tag{14}$$

If there exist tensors $\mathcal{W}^{\mathcal{L}}$ and $\mathcal{W}^{\mathcal{S}}$ satisfying these conditions, then the tensor $\mathcal{W}^\perp = \mathcal{W}^{\mathcal{L}} + \mathcal{W}^{\mathcal{S}}$ satisfies the conditions of equation (12), so exact recovery is certain. Similar to the argument in [12], we construct $\mathcal{W}^{\mathcal{L}}$ using a golfing scheme described in Section 4.1, and we construct $\mathcal{W}^{\mathcal{S}}$ as a the solution to a certain least-squares problem which is outlined in Section 4.2.

## 4.1 Constructing $\mathcal{W}^{\mathcal{L}}$

Our construction of $\mathcal{W}^{\mathcal{L}}$ uses a variation of the golfing scheme developed in [21, 20] and later used in [12, 13, 47]. Let $n \overset{\text{def}}{=} |\Omega^\perp| = d_1 d_2 d_3 - m = (1-\rho_s) d_1 d_2 d_3$. We create an i.i.d. uniformly distributed sequence of triples in $[d_1] \times [d_2] \times [d_3]$, call it $\{(a_i, b_i, c_i) : 1 \leq i \leq n\}$. This sequence is created by sampling with replacement from $\Omega^\perp$ using the following process:

1. Initialize $\mathfrak{S}_0 = \emptyset$.

2. For $i = 1, 2, \cdots, n$, sample the triple $(a_i, b_i, c_i)$ from $\mathfrak{S}_{i-1}$ uniformly with probability $|\mathfrak{S}_{i-1}|/d_1 d_2 d_3$, and sample $(a_i, b_i, c_i)$ uniformly from $\Omega^\perp \backslash \mathfrak{S}_{i-1}$ with probability $1 - |\mathfrak{S}_{i-1}|/d_1 d_2 d_3$.

3. Set $\mathfrak{S}_i = \mathfrak{S}_{i-1} \cup \{(a_i, b_i, c_i)\}$.

Here, each $\mathfrak{S}_i$ is a set containing triples corresponding to indices of the zero elements of $\mathcal{S}$. A similar scheme is used to construct a dual certificate for the tensor completion problem in [47] with an important distinction: we are sampling from $\Omega^\perp$ to analyze $\mathcal{S}$, while the scheme in [47] samples from $\Omega$. This is due to the fact that we observe samples in the complement of $\Omega$, while the opposite is true in the analysis of [47].

Notice that $\mathbb{P}((a_i, b_i, c_i) \in \mathfrak{S}_{i-1} | \mathfrak{S}_{i-1})$ is equal to the probability of the same event when the triples $(a_i, b_i, c_i)$ are drawn as i.i.d. random variables. Also, the conditional distribution of $(a_i, b_i, c_i)$ given $\mathfrak{S}_{i-1}$ and the event $(a_i, b_i, c_i) \in \mathfrak{S}_{i-1}^c$ is uniform. Together, these properties imply that the points $(a_i, b_i, c_i)$ are drawn uniformly from $[d_1] \times [d_2] \times [d_3]$ as i.i.d. random variables. Constructing this uniform sample via the golfing scheme is useful because it allows us to split our samples from $\Omega^\perp$ into a sequence of independent subsequences.

We split the sequence $\{(a_i, b_i, c_i) : 1 \leq i \leq n\}$ into $n_2$ subsets:

$$\Omega_k \overset{\text{def}}{=} \{(a_i, b_i, c_i) : n_1(k-1) \leq i \leq k n_1\},$$

where $|\Omega_k| \leq n_1$. Notice that $n_1 n_2 \leq n$ due to non-empty intersections among the $\Omega_k$. We choose the values $n_2 = \mathcal{O}(\log(n))$ and $n_1 = \mathcal{O}\left(\left(\frac{d_1 d_2 d_3 n}{\mu_0 \log(n)}\right)^{1/2}\right)$. The constants $n_1$ and $n_2$ must be chosen appropriately, and we show why we choose these particular values in later subsections.

With the sets $\Omega_k$ defined, we can define the corresponding projections $\mathcal{P}_{\Omega_k}$, and use these projections to construct $\mathcal{W}^{\mathcal{L}}$. With $\mathcal{Y}_0 = 0$, define the recursive sequence

$$\mathcal{Y}_j = \mathcal{Y}_{j-1} + \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\Omega_j} \mathcal{P}_{\mathcal{X}} (\mathcal{W} - \mathcal{Y}_{j-1}),$$

We set $\mathcal{W}^{\mathcal{L}} = \mathcal{P}_{\mathcal{X}^\perp}\mathcal{Y}_{n_2}$, and prove that this choice of $\mathcal{W}^{\mathcal{L}}$ satisfies conditions $(a), (b)$, and $(c)$ in (13) with high probability.

**Proof of** (13)$(\boldsymbol{a})$**.** Our argument uses ideas from [47]. In particular, we use the following two lemmas:

**Lemma 4** (Lem. 6 in [47])**.** *Let $\{(a_i, b_i, c_i)\}$ be an ordered set of independently and uniformly distributed samples from $[d_1] \times [d_2] \times [d_3]$ and $\Omega_j$ defined as above. Assume that $\mu(\mathcal{X}) \leq \mu_0$. Define $r \overset{def}{=} \overline{r}(\mathcal{X})$. Then for any fixed $j \in \{1, 2, \cdots, n_2\}$ and for all $\tau > 0$,*

$$\mathbb{P}\left( \left|\left|\left| \mathcal{P}_{\mathcal{X}} - \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\mathcal{X}} \mathcal{P}_{\Omega_j} \mathcal{P}_{\mathcal{X}} \right|\right|\right| \geq \tau \right) \tag{15}$$
$$\leq 2r^2(d_1 + d_2 + d_3) \ \exp\left( -\frac{n_1(\tau^2/2)}{(1 + 2\tau/3)(\mu_0^2 r^2 (d_1 + d_2 + d_3))} \right).$$

*Furthermore,*

$$\max_{\|\mathcal{T}\|_{\max}=1} \mathbb{P}\left( \left\| \left( \mathcal{P}_{\mathcal{X}} - \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\mathcal{X}} \mathcal{P}_{\Omega_j} \mathcal{P}_{\mathcal{X}} \right) \mathcal{T} \right\|_{\max} \geq \tau \right)$$
$$\leq 2 d_1 d_2 d_3 \exp\left( -\frac{n_1(\tau^2/2)}{(1 + 2\tau/3)\mu_0^2 r^2 (d_1 + d_2 + d_3)} \right).$$

**Lemma 5** (Lem. 7 in [47])**.** *Let $\alpha(\mathcal{X}) \leq \alpha_0$, $r \overset{def}{=} \overline{r}(\mathcal{X})$, and $q_1^* = (c + \log(d_1 + d_2 + d_3))^2 \alpha_0^2 r \log(d_1 + d_2 + d_3)$. There exists a positive constant $c_1$ so that for any constants $c > 0$ and $\delta_1 \in [1/(\log(d_1 + d_2 + d_3)), 1)$,*

$$n_1 \geq c_1 \left[ q_1^* (d_1 + d_2 + d_3)^{1+\delta_1} + \sqrt{q_1^*(1 + c)\delta_1^{-1} d_1 d_2 d_3} \right]$$

*implies*

$$\max_{\substack{\mathcal{P}_{\mathcal{X}}\mathcal{T} = \mathcal{T} \\ \|\mathcal{T}\|_{\max} \leq \|\mathcal{W}\|_{\max}}} \mathbb{P}\left( \left\| (\mathcal{P}_{\mathcal{X}} - \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\mathcal{X}} \mathcal{P}_{\Omega_1} \mathcal{P}_{\mathcal{X}}) \mathcal{T} \right\| \geq \frac{1}{16} \right) \leq (d_1 + d_2 + d_3)^{-c-1},$$

*where $\mathcal{W}$ is dual to $\mathcal{X}$.*

Lemma 5 puts a lower bound on acceptable choices for $n_1$, and because $n_1 n_2 \leq n$, this also puts an upper bound on acceptable choices for $n_2$. Our choices of $n_1$ and $n_2$ satisfy these conditions.

Instead of working directly with the sequence $\{\mathcal{Y}_j\}$, it is easier to work with the sequence

$$\mathcal{Z}_j \overset{\text{def}}{=} \mathcal{W} - \mathcal{P}_{\mathcal{X}} \mathcal{Y}_j.$$

Because $\mathcal{P}_{\mathcal{X}} \mathcal{W} = \mathcal{P}_{\mathcal{X}}^0 \mathcal{W} = \mathcal{W}$, it is clear that $\mathcal{P}_{\mathcal{X}} \mathcal{Z}_j = \mathcal{Z}_j$ for all $j$. From the definition of $\mathcal{Y}_j$, we derive the useful recursion

$$\mathcal{Z}_j = \mathcal{W} - \mathcal{P}_{\mathcal{X}} \left( \mathcal{Y}_{j-1} + \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\Omega_j} \mathcal{P}_{\mathcal{X}} (\mathcal{W} - \mathcal{P}_{\mathcal{X}} \mathcal{Y}_{j-1}) \right)$$

$$= \mathcal{P}_{\mathcal{X}} \left( I - \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\Omega_j} \right) \mathcal{P}_{\mathcal{X}} \mathcal{Z}_{j-1}. \tag{16}$$

We also see that

$$\begin{aligned} \mathcal{Y}_{n_2} &= \mathcal{Y}_{n_2 - 1} + \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\Omega_j} \mathcal{Z}_{n_2 - 1} \\ &= \mathcal{Y}_{n_2 - 2} + \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\Omega_j} \mathcal{Z}_{n_2 - 2} + \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\Omega_j} \mathcal{Z}_{n_2 - 1} \\ &= \sum_j \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\Omega_j} \mathcal{Z}_{j-1}. \end{aligned} \tag{17}$$

These facts imply

$$\begin{aligned} \mathcal{W}^{\mathcal{L}} &= \mathcal{P}_{\mathcal{X}^\perp} \mathcal{Y}_{n_2} \\ &= \mathcal{P}_{\mathcal{X}^\perp} \sum_j \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\Omega_j} \mathcal{Z}_{j-1} \\ &= \mathcal{P}_{\mathcal{X}^\perp} \sum_j \left( \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\Omega_j} - I \right) \mathcal{Z}_{j-1}, \end{aligned} \tag{18}$$

where the last equality follows from the fact that $\mathcal{P}_{\mathcal{X}} \mathcal{Z}_{j-1} = \mathcal{Z}_{j-1}$. We now use the sequence $\{\mathcal{Z}_j\}$ to show that $\|\mathcal{W}^{\mathcal{L}}\| < \frac{1}{8}$. For convenience, let $\mathcal{R}_j = I - \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\Omega_j}$. To prove the bound $\|\mathcal{W}^{\mathcal{L}}\| < \frac{1}{8}$, we decompose $\mathcal{W}^{\mathcal{L}}$ into the sum shown in equation (18) and use Lemma 4 to bound the spectral norm of each of the terms.

$$\begin{aligned} \mathbb{P}\left( \|\mathcal{P}_{\mathcal{X}^\perp} \mathcal{Y}_{n_2}\| \geq \frac{1}{8} \right) &\leq \mathbb{P}\left( \left\| \sum_{j=1}^{n_2} \mathcal{R}_j \mathcal{Z}_{j-1} \right\| \geq \frac{1}{8} \right) \\ &\leq \mathbb{P}\left( \|\mathcal{R}_1 \mathcal{Z}_0\| \geq \frac{1}{16} \right) + \mathbb{P}\left( \|\mathcal{Z}_1\|_{\max} \geq \|\mathcal{W}\|_{\max}/4 \right) \\ &\quad + \mathbb{P}\left( \left\| \sum_{j=2}^{n_2} \mathcal{R}_j \mathcal{Z}_{j-1} \right\| \geq \frac{1}{16},\ \|\mathcal{Z}_1\|_{\max} < \|\mathcal{W}\|_{\max}/4 \right) \\ &\leq \mathbb{P}\left( \|\mathcal{R}_1 \mathcal{Z}_0\| \geq \frac{1}{16} \right) + \mathbb{P}\left( \|\mathcal{Z}_1\|_{\max} \geq \|\mathcal{W}\|_{\max}/4 \right) \\ &\quad + \mathbb{P}\left( \|\mathcal{R}_2 \mathcal{Z}_1\| \geq \frac{1}{32},\ \|\mathcal{Z}_1\|_{\max} < \|\mathcal{W}\|_{\max}/4 \right) \\ &\quad + \mathbb{P}\left( \|\mathcal{Z}_2\|_{\max} \geq \|\mathcal{W}\|_{\max}/8, \|\mathcal{Z}_1\|_{\max} < \|\mathcal{W}\|_{\max}/4 \right) \\ &\quad + \mathbb{P}\left( \left\| \sum_{j=3}^{n_2} \mathcal{R}_j \mathcal{Z}_{j-1} \right\| \geq \frac{1}{32}, \|\mathcal{Z}_2\|_{\max} < \|\mathcal{W}\|_{\max}/8 \right) \\ &\leq \mathbb{P}\left( \|\mathcal{R}_1 \mathcal{Z}_0\| \geq \frac{1}{16} \right) + \mathbb{P}\left( \|\mathcal{Z}_1\|_{\max} \geq \|\mathcal{W}\|_{\max}/4 \right) \\ &\quad + \sum_{j=2}^{n_2 - 1} \mathbb{P}\big( \|\mathcal{P}_{\mathcal{X}^*} \mathcal{R}_j \mathcal{P}_{\mathcal{X}^*} \mathcal{Z}_{j-1}\|_{\max} \geq \|\mathcal{W}\|_{\max}/2^{j+1}, \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad \|\mathcal{Z}_{j-1}\|_{\max} \leq \|\mathcal{W}\|_{\max}/2^j\big) \\
&\quad + \sum_{j=2}^{n_2} \mathbb{P}\left(\|\mathcal{R}_j \mathcal{Z}_{j-1}\| \geq 2^{-3-j},\ \|\mathcal{Z}_{j-1}\|_{\max} \leq \|\mathcal{W}\|_{\max}/2^j\right) \\
&\leq \mathbb{P}\left(\|\mathcal{R}_1 \mathcal{Z}_0\| \geq \frac{1}{16}\right) + \mathbb{P}\left(\|\mathcal{Z}_1\|_{\max} \geq \|\mathcal{W}\|_{\max}/4\right) \\
&\quad + \sum_{j=2}^{n_2-1} \mathbb{P}\left(\|\mathcal{P}_{\mathcal{X}^*} \mathcal{R}_j \mathcal{P}_{\mathcal{X}^*} \mathcal{Z}_{j-1}\|_{\max} \geq \|\mathcal{Z}_{j-1}\|_{\max}/2\right) \\
&\quad + \sum_{j=2}^{n_2} \mathbb{P}\left(\|\mathcal{R}_j \mathcal{Z}_{j-1}\| \geq 2^{-3}\|\mathcal{Z}_{j-1}\|_{\max}/\|\mathcal{W}\|_{\max}\right).
\end{aligned}
$$

For the penultimate inequality, we applied the recursion (16). Because $(\mathcal{R}_j,\ \mathcal{Z}_{j-1})$ are independent, Lemma 4 with the maximizing $\mathcal{T} = \mathcal{Z}_{j-1}/\|\mathcal{Z}_{j-1}\|_{\max}$ and $\tau = \frac{1}{8}$ gives the bound

$$
\begin{aligned}
\mathbb{P}\left(\|\mathcal{P}_{\mathcal{X}^\perp} \mathcal{Y}_{n_2}\| \geq \frac{1}{8}\right) &\leq \max_{\substack{\mathcal{P}_{\mathcal{X}}\mathcal{T}=\mathcal{T} \\ \|\mathcal{T}\|_{\max}\leq 1}} \left[\sum_{j=1}^{n_2-1} \mathbb{P}\left(\|\mathcal{P}_{\mathcal{X}} \mathcal{R}_j \mathcal{P}_{\mathcal{X}} \mathcal{T}\|_{\max} \geq \frac{1}{4}\right)\right. \\
&\qquad \left. + \sum_{j=1}^{n_2} \mathbb{P}\left(\|\mathcal{R}_j \mathcal{T}\| \geq \frac{1}{16\|\mathcal{W}\|_{\max}}\right)\right] \\
&\leq n_2 \left[\max_{\substack{\mathcal{P}_{\mathcal{X}}\mathcal{T}=\mathcal{T} \\ \|\mathcal{T}\|_{\max}\leq 1}} \mathbb{P}\left(\|\mathcal{P}_{\mathcal{X}} \mathcal{R}_1 \mathcal{P}_{\mathcal{X}} \mathcal{T}\|_{\max} > \frac{1}{4}\right)\right. \\
&\qquad \left. + \mathbb{P}\left(\|\mathcal{R}_1 \mathcal{T}\| > \frac{1}{16\|\mathcal{W}\|_{\max}}\right)\right] \\
&\leq 2n_2 d_1 d_2 d_3 \exp\left(-\frac{(3/112)n_1}{\mu_0^2 r^2 (d_1+d_2+d_3)}\right) \\
&\qquad + n_2 \left[\max_{\substack{\mathcal{P}_{\mathcal{X}}\mathcal{T}=\mathcal{T} \\ \|\mathcal{T}\|_{\max}\leq \|\mathcal{W}\|_{\max}}} \mathbb{P}\left(\|\mathcal{R}_1 \mathcal{T}\| > \frac{1}{16}\right)\right].
\end{aligned}
$$

Applying Lemma 5 to bound the final term, we have shown

$$
\begin{aligned}
&\mathbb{P}\Big(\|\mathcal{P}_{\mathcal{X}^\perp} \mathcal{Y}_{n_2}\| \geq \frac{1}{8}\Big) \\
&\qquad\qquad \leq 2n_2 d_1 d_2 d_3 \exp\left(-\frac{(3/112)n_1}{\mu_0^2 r^2 (d_1+d_2+d_3)}\right) + (d_1+d_2+d_3)^{-1-c}.
\end{aligned}
$$

This shows that $\mathcal{W}^{\mathcal{L}} = \mathcal{P}_{\mathcal{X}^\perp} \mathcal{Y}_{n_2}$ satisfies condition (13)(a) with high probability.

**Proof of** (13)(***b***)**.** We would like to prove that $\|\mathcal{P}_\Omega(\mathcal{W} + \mathcal{W}^{\mathcal{L}})\|_F < \frac{\lambda}{8}$. By the definition of the operator norm in (3), equation (15) of Lemma 4 implies

$$
\left\|\left(\mathcal{P}_{\mathcal{X}} - \frac{d_1 d_2 d_3}{n_1} \mathcal{P}_{\mathcal{X}} \mathcal{P}_{\Omega_j} \mathcal{P}_{\mathcal{X}}\right) \mathcal{T}\right\|_F \leq \tau \|\mathcal{T}\|_F
$$

with high probability for all $\mathcal{T}$ satisfying $\mathcal{P}_{\mathcal{X}}\mathcal{T} = \mathcal{T}$. Consequentially, using the independence of $\Omega_j$ and $\mathcal{Z}_{j-1}$, we have

$$\forall j, \quad \|\mathcal{Z}_j\|_F \leq \tau \|\mathcal{Z}_{j-1}\|_F \implies \|\mathcal{Z}_{n_2}\|_F \leq \tau^{n_2} \|\mathcal{W}\|_F.$$

Using equation (2) and assuming the coherence of $\mathcal{X}$ satisfies $\alpha(\mathcal{X}) \leq \alpha_0$,

$$\|\mathcal{W}\|_{\max} \leq \alpha_0 \left( \frac{\overline{r}}{d_1 d_2 d_3} \right)^{1/2}.$$

Therefore, since $\|\mathcal{X}\|_F \leq \sqrt{d_1 d_2 d_3} \|\mathcal{X}\|_{\max}$ for all tensors $\mathcal{X}$ of size $d_1 \times d_2 \times d_3$,

$$\|\mathcal{Z}_{n_2}\|_F \leq \tau^{n_2} \alpha_0 \sqrt{\overline{r}}.$$

Hence,

$$\begin{aligned} \|\mathcal{P}_\Omega(\mathcal{W} + \mathcal{W}^{\mathcal{L}})\|_F &= \|\mathcal{P}_\Omega(\mathcal{W} + (I - \mathcal{P}_{\mathcal{X}})\mathcal{Y}_{n_2})\|_F \\ &= \|\mathcal{P}_\Omega \mathcal{Z}_{n_2}\|_F \\ &\leq \tau^{n_2} \alpha_0 \sqrt{\overline{r}}, \end{aligned}$$

where we used the fact that $\mathcal{P}_\Omega \mathcal{Y}_{n_2} = 0$. Let $\tau = \mathcal{O}(e^{-1})$ and $n_2 = \mathcal{O}(\log(n))$. Because $\overline{r} \leq \rho_r \left( \frac{n}{(d_1+d_2+d_3)\log(n)\alpha_0^4 \mu_0^2} \right)^{1/2}$ (cf. (7)), the above bound is smaller than $\frac{\lambda}{8}$ as long as $\rho_r$ is a small enough constant.

These parameter choices also ensure the probability that this bound holds, given in Lemma 5, is large. For the sequel, we require that $\|\mathcal{P}_\Omega(\mathcal{W} + \mathcal{W}^{\mathcal{L}})\|_F < \frac{\lambda}{16}$, and it is clear that this bound also holds with high probability.

**Proof of** (13)**(*c*).** We would like to prove that $\|\mathcal{P}_{\Omega^\perp}(\mathcal{W} + \mathcal{W}^{\mathcal{L}})\|_{\max} < \frac{\lambda}{8}$. We have that $\mathcal{W} + \mathcal{W}^{\mathcal{L}} = \mathcal{Y}_{n_2} + \mathcal{Z}_{n_2}$, so $\|\mathcal{P}_{\Omega^\perp}(\mathcal{W} + \mathcal{W}^{\mathcal{L}})\|_{\max} \leq \|\mathcal{Y}_{n_2}\|_{\max} + \|\mathcal{Z}_{n_2}\|_{\max}$. From the previous section, we already have the bound $\|\mathcal{Z}_{n_2}\|_{\max} \leq \|\mathcal{Z}_{n_2}\|_F \leq \frac{\lambda}{16}$, so we must only bound $\|\mathcal{Y}_{n_2}\|_{\max}$.

$$\begin{aligned} \|\mathcal{Y}_{n_2}\|_{\max} &= \left( \frac{d_1 d_2 d_3}{n_1} \right) \left\| \sum_{j=1}^{n_2} \mathcal{P}_{\Omega_j} \mathcal{Z}_{j-1} \right\|_{\max} && \text{(By equation (17))} \\ &\leq \left( \frac{d_1 d_2 d_3}{n_1} \right) \sum_{j=1}^{n_2} \left\| \mathcal{P}_{\Omega_j} \mathcal{Z}_{j-1} \right\|_{\max} \\ &\leq \left( \frac{d_1 d_2 d_3}{n_1} \right) \sum_{j=1}^{n_2} \|\mathcal{Z}_{j-1}\|_{\max} \\ &\leq \left( \frac{d_1 d_2 d_3}{n_1} \right) \left( \sum_{j=0}^{n_2-1} \tau^j \right) \|\mathcal{W}\|_{\max} \\ &\leq \left( \frac{d_1 d_2 d_3}{n_1} \right) \left( \sum_{j=0}^{n_2-1} \tau^j \right) \alpha_0 \left( \frac{\overline{r}}{d_1 d_2 d_3} \right)^{1/2} && \text{(By equation (2))} \end{aligned}$$

$$\leq \left(\frac{\alpha_0 \sqrt{\overline{r}} d_1 d_2 d_3}{n_1}\right)(1-\tau)^{-1}.$$

With $n_1 = \mathcal{O}\left(\left(\frac{d_1 d_2 d_3 n}{\mu_0 \log(n)^2}\right)^{1/2}\right)$ and $\overline{r} \leq \rho_r \left(\frac{n}{(d_1+d_2+d_3)\log(n)\alpha_0^4 \mu_0^2}\right)^{1/2}$, it is clear that $\|\mathcal{Y}_{n_2}\| \leq \frac{\lambda}{16}$ when $\rho_r$ and $\rho_s$ are small enough, so the desired result holds.

These choices of parameters are consistent; it is straightforward to see that $n_1 n_2 \leq n$, and that these choices of $n_1$ and $\tau$ ensure that the bounds outlined in Lemma 4 hold with exponentially small probability.

## 4.2 Constructing $\mathcal{W}^{\mathcal{S}}$

In the previous sections, we supposed that the support of $\Omega$ was uniformly distributed over all sets of cardinality $m$. For our construction of $\mathcal{W}^{\mathcal{S}}$, it is easier to work under the assumption that the support of $\Omega$ follows a Bernoulli distribution with parameter $\rho_s$:

$$\Omega_{i,j,k} = \begin{cases} 1 & \text{w.p. } \rho_s, \\ 0 & \text{w.p. } 1-\rho_s. \end{cases} \tag{19}$$

We can construct $\mathcal{W}^{\mathcal{S}}$ under this model and show that $\mathcal{W}^{\mathcal{S}}$ satisfies conditions (14)($d$) and (14)($e$) with high probability. It follows that $\mathcal{W}^{\mathcal{S}}$ under the original uniform model satisfies conditions (14)($d$) and (14)($e$) with high probability as well. This correspondence between the Bernoulli and uniform distributions is well-known and used in [12] as well, although in a slightly different way. We include a proof in Appendix B for completeness.

We make one more simplification to our model before constructing $\mathcal{W}^{\mathcal{S}}$. Under the model of Theorem 1, the signs of the entries of $\mathcal{S}$ are arbitrary, but it is more convenient to assume that the signs follow a Bernoulli model with parameter $\frac{\rho_s}{2}$. The equivalence of these two models is also well-known and used in [12]. We save a formal discussion of this equivalence for Appendix B. With these changes to our model in place, we are now prepared to construct $\mathcal{W}^{\mathcal{S}}$.

Following the ideas behind the construction of the dual certificate for matrix RPCA [12], we let

$$\mathcal{W}^{\mathcal{S}} = \lambda \mathcal{P}_{\mathcal{X}^\perp}(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_{\mathcal{X}} \mathcal{P}_\Omega)^{-1} \text{sgn}(\mathcal{S}).$$

Our assumption $|||\mathcal{P}_\Omega \mathcal{P}_{\mathcal{X}}||| < \frac{1}{2}$ implies $|||\mathcal{P}_{\mathcal{X}} \mathcal{P}_\Omega \mathcal{P}_{\mathcal{X}}||| < \frac{1}{4}$ since these are orthogonal projections, so the inverse of $\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_{\mathcal{X}} \mathcal{P}_\Omega$ as an operator mapping the range of $\mathcal{P}_\Omega$ onto itself exists. It is also important that $\mathcal{P}_\Omega \mathcal{W}^{\mathcal{S}} = \lambda \mathcal{P}_\Omega (I - \mathcal{P}_{\mathcal{X}})(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_{\mathcal{X}} \mathcal{P}_\Omega)^{-1} \text{sgn}(\mathcal{S}) = \lambda \text{sgn}(\mathcal{S})$. As in the matrix case, $\mathcal{W}^{\mathcal{S}}$ can be interpreted as the tensor with minimum Frobenius norm over the set $\{\mathcal{T} : \mathcal{P}_{\mathcal{X}^\perp} \mathcal{T} = \mathcal{T},\ \mathcal{P}_\Omega \mathcal{T} = \lambda \text{sgn}(\mathcal{S})\}$.

**Proof of** (14)**(*d*).** With $\mathcal{W}^{\mathcal{S}} = \lambda \mathcal{P}_{\mathcal{X}^\perp}(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_{\mathcal{X}} \mathcal{P}_\Omega)^{-1} \text{sgn}(\mathcal{S})$, we would like to show that $\|\mathcal{W}^{\mathcal{S}}\| < \frac{1}{8}$. Let $G = \text{sgn}(\mathcal{S})$ for convenience. The elements of $G$ follow the distribution

$$G_{i,j,k} = \begin{cases} 1 & \text{with probability } \frac{\rho_s}{2}, \\ 0 & \text{with probability } 1-\rho_s, \\ -1 & \text{with probability } \frac{\rho_s}{2}. \end{cases}$$

We can then write

$$\begin{aligned}
\|\mathcal{W}^{\mathcal{S}}\| &= \lambda \left\|\mathcal{P}_{\mathcal{X}^\perp}(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega)^{-1}G\right\| \\
&= \lambda \left\|\mathcal{P}_{\mathcal{X}^\perp}\sum_{k=0}^{\infty}(\mathcal{P}_\Omega\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega)^k G\right\| \\
&\leq \lambda \left\|\mathcal{P}_{\mathcal{X}^\perp} G\right\| + \lambda \left\|\mathcal{P}_{\mathcal{X}^\perp}\sum_{k=1}^{\infty}(\mathcal{P}_\Omega\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega)^k G\right\| \\
&\leq \lambda \left\|G\right\| + \lambda \left\|\sum_{k=1}^{\infty}(\mathcal{P}_\Omega\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega)^k G\right\|,
\end{aligned}$$

where we have used the Neumann series expansion of the operator $(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega)^{-1}$. The first term can be bounded using existing tail bounds on the spectral norm of random tensors. The distribution of $G$ is subgaussian, so we can apply the following result from [43].

**Lemma 6.** *The following holds with probability at least* $1 - \delta$*:*

$$\|G\| \leq \sqrt{8\,(d_1 + d_2 + d_3)\log(6/\log(3/2)) + \log(2/\delta)}$$

*Proof.* This result is an application of [43, Thm. 1]. See Appendix C for the details. □

With $\lambda = (d_1 + d_2 + d_3)^{-\frac{1}{2}}$, Lemma 6 ensures that $\lambda\|G\| \leq \frac{1}{16}$ with large probability. To bound the second term, we use an $\epsilon$-net covering argument. Define the following set of "digitalized" vectors:

$$\mathcal{B}_{m_j,d_j} = \{0, \pm 1, \pm 2^{-1/2}, \cdots, \pm 2^{-m_j/2}\}^{d_j} \cap \{u \in \mathbb{R}^{d_j} : \|u\| \leq 1\}.$$

Let $\mathcal{Q}$ be the operator $\sum_{k=1}^{\infty}(\mathcal{P}_\Omega\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega)^k$. We can bound $\|\mathcal{Q}(G)\|$ by considering the action of $\mathcal{Q}$ on tensor product spaces of the $\mathcal{B}_{m_j,d_j}$. With $m_j = \lceil \log_2(d_j) \rceil$, the following bound holds by Lemma 9 in [47]:

$$\|\mathcal{Q}(G)\| = \max_{u_j \in \mathbb{S}^{d_j}} \langle \mathcal{Q}(G), u_1 \otimes u_2 \otimes u_3 \rangle \leq 8 \max_{u_j \in \mathcal{B}_{m_j,d_j}} \langle \mathcal{Q}(G), u_1 \otimes u_2 \otimes u_3 \rangle, \tag{20}$$

where $\mathbb{S}^{d_j}$ is the unit sphere of dimension $d_j$. Because $Q$ is self-adjoint, this also implies

$$\|\mathcal{Q}(G)\| = \max_{u_j \in \mathcal{S}^{d_j}} \big\langle \mathcal{Q}(G), u_1 \otimes u_2 \otimes u_3 \big\rangle \leq 8 \max_{u_j \in \mathcal{B}_{m_j,d_j}} \langle G, \mathcal{Q}(u_1 \otimes u_2 \otimes u_3) \rangle.$$

Let $X(u,v,w) \overset{\text{def}}{=} \big\langle G, Q(u_1 \otimes u_2 \otimes u_3) \big\rangle$ for convenience. Because the signs of $G$ are i.i.d. symmetric, we can apply Hoeffding's inequality, conditional on the event that the support of $G$ is exactly $\Omega$:

$$\mathbb{P}(|X(u_1,u_2,u_3)| > t \mid \Omega) \leq 2\exp\left(-\frac{2t^2}{\|\mathcal{Q}(u_1 \otimes u_2 \otimes u_3)\|_F^2}\right).$$

Because $\|\mathcal{Q}(u_1 \otimes u_2 \otimes u_3)\|_F^2 \leq |||\mathcal{Q}|||^2 \cdot \|(u_1 \otimes u_2 \otimes u_3)\|_F^2 = |||\mathcal{Q}|||^2$,

$$\mathbb{P}(|X(u_1,u_2,u_3)| > t \mid \Omega) \leq 2\exp\left(-\frac{2t^2}{|||Q|||^2}\right).$$

Taking the maximum of $|X(u_1, u_2, u_3)|$ with $u_j \in \mathcal{B}_{m_j,d_j},\ j = 1, 2, 3$, we have

$$\mathbb{P}\left(\max_{u_j \in \mathcal{B}_{m_j,d_j}} |X(u_1, u_2, u_3)| > t \;\middle|\; \Omega\right) \leq 2\left(\prod_{j=1,2,3} |\mathcal{B}_{m_j,d_j}|\right) \exp\left(-\frac{2t^2}{\|\!|\mathcal{Q}|\!\|^2}\right).$$

Applying (20) to this inequality yields

$$\mathbb{P}(\|\mathcal{Q}(G)\| > t \mid \Omega) \leq \left(\prod_{j=1,2,3} |\mathcal{B}_{m_j,d_j}|\right) \exp\left(-\frac{t^2}{32\|\!|\mathcal{Q}|\!\|^2}\right).$$

We can bound $\|\!|\mathcal{Q}|\!\|$ conditional on the event that $\|\!|\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}|\!\| \leq \sigma$. (See Section 4.3 for a proof that this event holds with high probability.) Recall $\|\!|\mathcal{Q}|\!\| = \left\|\!\left|\sum_{k=1}^\infty (\mathcal{P}_\Omega \mathcal{P}_\mathcal{X} \mathcal{P}_\Omega)^k\right|\!\right\| \leq \sum_{k=1}^\infty \left\|\!\left|((\mathcal{P}_\Omega \mathcal{P}_\mathcal{X})(\mathcal{P}_\Omega \mathcal{P}_\mathcal{X})^*)^k\right|\!\right\| \leq \frac{\sigma^2}{1-\sigma^2}$. We can also bound the cardinality of $\mathcal{B}_{m_j,d_j}$ using equation (21) of [47]:

$$\prod_{j=1,2,3} |\mathcal{B}_{m_j,d_j}| = e^{(21/4)(d_1+d_2+d_3)}.$$

This gives us the unconditional bound

$$\begin{aligned}
&\mathbb{P}(\lambda\|Q(G)\| > t) \qquad (21)\\
&\leq 2\left(\prod_{j=1,2,3} |\mathcal{B}_{m_j,d_j}|\right) \exp\left(-\frac{t^2(1-\sigma^2)^2}{32\sigma^4\lambda^2}\right) + \mathbb{P}(\|\!|\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}|\!\| > \sigma)\\
&\leq 2\left(\exp\left(\frac{21(d_1+d_2+d_3)}{4}\right)\right) \exp\left(-\frac{t^2(1-\sigma^2)^2}{32\sigma^4\lambda^2}\right) + \mathbb{P}(\|\!|\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}|\!\| > \sigma)\\
&= 2\exp\left(-\frac{t^2(1-\sigma^2)^2}{32\sigma^4\lambda^2} + \frac{21(d_1+d_2+d_3)}{4}\right) + \mathbb{P}(\|\!|\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}|\!\| > \sigma).
\end{aligned}$$

This shows that if $\lambda$ is chosen to be on the order of $(d_1+d_2+d_3)^{-1/2}$, and if $\sigma$ is a small-enough constant, then $\|\mathcal{W}^\mathcal{S}\| < \frac{1}{8}$ with high probability.

**Proof of** (14)***(e)*****.** We would like to show that $\|\mathcal{P}_{\Omega^\perp}\mathcal{W}^\mathcal{S}\|_{\max} < \frac{\lambda}{8}$. By our definition of $\mathcal{W}^\mathcal{S}$,

$$\begin{aligned}
\mathcal{W}^\mathcal{S} &= \lambda \mathcal{P}_{\mathcal{X}^\perp}(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\mathcal{X} \mathcal{P}_\Omega)^{-1} G\\
&= \lambda(I - \mathcal{P}_\mathcal{X})(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\mathcal{X} \mathcal{P}_\Omega)^{-1} G\\
&= -\lambda \mathcal{P}_\mathcal{X}(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\mathcal{X} \mathcal{P}_\Omega)^{-1} G
\end{aligned}$$

Choosing $(i, j, k) \in \Omega^\perp$, we have

$$\begin{aligned}
\mathcal{W}^\mathcal{S}_{i,j,k} &= \left\langle \mathcal{W}^\mathcal{S}, e_i \otimes e_j \otimes e_k \right\rangle\\
&= -\lambda \left\langle \mathcal{P}_\mathcal{X}(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\mathcal{X} \mathcal{P}_\Omega)^{-1} G, e_i \otimes e_j \otimes e_k \right\rangle\\
&= -\lambda \left\langle \mathcal{P}_\mathcal{X}(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\mathcal{X} \mathcal{P}_\Omega)^{-1} G, e_i \otimes e_j \otimes e_k \right\rangle\\
&= -\lambda \left\langle G, (\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\mathcal{X} \mathcal{P}_\Omega)^{-1} \mathcal{P}_\Omega \mathcal{P}_\mathcal{X}(e_i \otimes e_j \otimes e_k) \right\rangle,
\end{aligned}$$

where we used the fact that $\mathcal{P}_\Omega,\ \mathcal{P}_\mathcal{X}$, and $(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega)^{-1}$ are self-adjoint. For convenience, let

$$R(i,j,k) = (\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega)^{-1}\mathcal{P}_\Omega\mathcal{P}_\mathcal{X}(e_i \otimes e_j \otimes e_k).$$

Now we bound the maximum entry of $\mathcal{W}^\mathcal{S}$ with high probability conditional on the events that the support of $G$ is exactly $\Omega$ and that $\lVert\!|\mathcal{P}_\Omega\mathcal{P}_\mathcal{X}|\!\rVert \le \sigma$. Because the entries of $G$ are i.i.d. symmetric, Hoeffding's inequality gives

$$\mathbb{P}\left(\left|\mathcal{W}^\mathcal{S}_{i,j,k}\right| \ge \eta \mid \Omega\right) \le 2\exp\left(-\frac{2\eta^2}{\lambda^2\|R(i,j,k)\|_F^2}\right),$$

so

$$\mathbb{P}\left(\max_{i,j,k}\left|\mathcal{W}^\mathcal{S}_{i,j,k}\right| \ge \eta \mid \Omega\right) \le 2d_1d_2d_3\exp\left(-\frac{2\eta^2}{\lambda^2\max\limits_{i,j,k}\|R(i,j,k)\|_F^2}\right).$$

All that is left is to bound $\|R(i,j,k)\|_F^2$. We need the following lemma from [47]:

**Lemma 7** (Lem. 2 in [47])**.** *Let $\mathcal{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$ be a third-order tensor. Then*

$$\max_{i,j,k}\|\mathcal{P}_\mathcal{X}(e_i \otimes e_j \otimes e_k)\|_F^2 \le \frac{\overline{r}(\mathcal{X})^2(d_1+d_2+d_3)}{d_1d_2d_3}\mu(\mathcal{X})^2.$$

By Lemma 7,

$$\|\mathcal{P}_\mathcal{X}(e_i \otimes e_j \otimes e_k)\|_F \le \mu_0\overline{r}\left(\frac{d_1+d_2+d_3}{d_1d_2d_3}\right)^{1/2}.$$

Hence,

$$\begin{aligned}\|\mathcal{P}_\Omega\mathcal{P}_\mathcal{X}(e_i \otimes e_j \otimes e_k)\|_F &\le (\lVert\!|\mathcal{P}_\Omega\mathcal{P}_\mathcal{X}|\!\rVert)(\|\mathcal{P}_\mathcal{X}(e_i \otimes e_j \otimes e_k)\|_F)\\ &\le \sigma\mu_0\overline{r}\left(\frac{d_1+d_2+d_3}{d_1d_2d_3}\right)^{1/2}\end{aligned}$$

Furthermore,

$$\begin{aligned}\|(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega)^{-1}\| &= \|(\mathcal{P}_\Omega - (\mathcal{P}_\Omega\mathcal{P}_\mathcal{X})(\mathcal{P}_\Omega\mathcal{P}_\mathcal{X})^*)^{-1}\|\\ &\le (1-\sigma^2)^{-1}.\end{aligned}$$

Combining these two bounds, we have

$$\begin{aligned}\|R(i,j,k)\|_F^2 &= \|(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega)^{-1}\mathcal{P}_\Omega\mathcal{P}_\mathcal{X}(e_i \otimes e_j \otimes e_k)\|_F^2\\ &\le \left(\frac{\sigma\mu_0\overline{r}}{1-\sigma^2}\right)^2\left(\frac{d_1+d_2+d_3}{d_1d_2d_3}\right).\end{aligned}$$

Finally, we have derived that

$$\begin{aligned}\mathbb{P}\left(\max_{i,j,k}\left|\mathcal{W}^\mathcal{S}_{i,j,k}\right| \ge \eta\right) \le& 2d_1d_2d_3\exp\left(-\frac{2\eta^2(1-\sigma^2)^2d_1d_2d_3}{(\sigma\lambda\mu_0\overline{r})^2(d_1+d_2+d_3)}\right)\\ &+\mathbb{P}\left(\lVert\!|\mathcal{P}_\Omega\mathcal{P}_\mathcal{X}|\!\rVert > \sigma\right).\end{aligned} \tag{22}$$

Letting $\eta = \frac{\lambda}{8}$, this shows that if $\overline{r} \le \rho_r\left(\frac{n}{(d_1+d_2+d_3)\log(n)\alpha_0^4\mu_0^2}\right)^{1/2}$, then $\|\mathcal{P}_{\Omega^\perp}\mathcal{W}^\mathcal{S}\|_{\max} < \frac{\lambda}{8}$ with high probability, provided that $\rho_r$ is a small-enough constant.

### 4.3 Bounding the Operator Norm of $\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}$

The past two results were conditional on the event that $|||\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}||| \le \sigma$ with high probability. Lemma 5 from [47] shows this is true.

**Lemma 8** (Lem. 5 in [47])**.** *Let $\mu(\mathcal{X}) \le \mu_0$, $\overline{r}(\mathcal{X}) = r$, and $\Omega$ follow the Bernoulli model of equation* (19)*. Then, for any $\tau > 0$*

$$\mathbb{P}\Big( |||\mathcal{P}_\mathcal{X}(\tfrac{d_1 d_2 d_3}{n}\mathcal{P}_\Omega - I)\mathcal{P}_\mathcal{X}||| \ge \sigma \Big)$$
$$\le 2r^2(d_1 + d_2 + d_3)\exp\left( -\frac{n\sigma^2/2}{(1 + 2\sigma/3)\mu_0^2 r^2 (d_1 + d_2 + d_3)} \right).$$

By our assumption in (7), $r \le \mathcal{O}\left( \left( \frac{n}{(d_1+d_2+d_3)\log(n)\alpha_0^4 \mu_0^2} \right)^{\frac{1}{2}} \right)$, so as long as $n$ is not too large (or, equivalently, $\rho_s$ is sufficiently small) Lemma 8 shows that as an operator in the range of $\mathcal{P}_\mathcal{X}$, $\frac{d_1 d_2 d_3}{n} |||\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}||| \in [1/2, 3/2]$. Using the fact that $|||\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}|||^2 = |||(\mathcal{P}_\Omega \mathcal{P}_\mathcal{X})^*(\mathcal{P}_\Omega \mathcal{P}_\mathcal{X})||| = |||\mathcal{P}_\mathcal{X}\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}|||$, we have also bounded $|||\mathcal{P}_\Omega \mathcal{P}_\mathcal{X}|||$ with high probability.

Lemma 8 assumes the support of $\Omega$ is uniformly distributed over all sets of cardinality $m$, but in Section 4.2, we assume that $\Omega$ follows the Bernoulli model. In Appendix B, we show that these two models are essentially equivalent, so the conclusion of Lemma 8 holds for the Bernoulli model as well.

## 5 A Nonconvex Approach to Atomic Norm Minimization

Although Theorem 1 shows that the program (6) can exactly recover a low Tucker-rank tensor and a sparse tensor from their superposition, (6) is NP-hard to solve in general due to the intractability of the atomic norm [23]. For low-rank matrix recovery, it is common to accelerate computation by replacing the nuclear norm with a nonconvex formulation based on a factorization of $X$ [13, 16, 40],

$$\|X\|_* = \inf_{UV^T = X} \quad \tfrac{1}{2}(\|U\|_F^2 + \|V\|_F^2). \tag{23}$$

A similar factorized formulation of the tensor atomic norm has been used for order-three tensors in [9] and [42] for tensor completion. Due to the non-smooth regularizer in the tensor RPCA problem, this factorization approach must be handled with care. We introduce our factorized program in section 5.1 and discuss a particularly fast method for handling the non-smooth regularizer in section 5.2. We conclude this section by showing that many local minima of our factorized program are globally optimal.

### 5.1 Burer-Monteiro Factorization in Higher-Orders

We now explicitly work with tensors of order-$K$, with $K \ge 3$. Instead of establishing equivalence to the program (6) directly, we work with its Lagrangian formulation:

$$\min_{\mathcal{X},\mathcal{S}} \quad \frac{1}{2}\|\mathcal{X} + \mathcal{S} - \mathcal{Z}\|_F^2 + \lambda_x \|\mathcal{X}\|_* + \lambda_s \|\mathcal{S}\|_{\text{sum}}. \tag{24}$$

A solution $(\mathcal{X}^*, \mathcal{S}^*)$ to (24) solves a variant of (6) where the equality constraint is replaced with $\|\mathcal{X} + \mathcal{S} - \mathcal{Z}\|_F \leq \epsilon$ with $\epsilon \stackrel{\text{def}}{=} \|\mathcal{X}^* + \mathcal{S}^* - \mathcal{Z}\|_F$ and $\lambda = \lambda_s/\lambda_x$.

Our factorized approach implicitly introduces a bound on the rank of $\mathcal{X}$, yielding the constrained problem

$$\min_{\mathcal{X},\mathcal{S}} \quad \frac{1}{2}\|\mathcal{X} + \mathcal{S} - \mathcal{Z}\|_F^2 + \lambda_x\|\mathcal{X}\|_* + \lambda_s\|\mathcal{S}\|_{\text{sum}} \quad \text{s.t.} \quad \text{rank}_{\text{atomic}}(\mathcal{X}) \leq R. \tag{25}$$

Although (25) is nonconvex, it has the same global optima as the convex program (24) as long as the rank-bound $R$ is non-restrictive at the solution [45]. We can explicitly parameterize the nonconvex model (25) as

$$\min_{\{a_r^{(1)}\},\cdots,\{a_r^{(K)}\},\mathcal{S}} \quad \frac{1}{2}\left\| \sum_{r=1}^{R}(a_r^{(1)} \otimes \cdots \otimes a_r^{(K)}) + \mathcal{S} - \mathcal{Z} \right\|_F^2 + \lambda_s\|\mathcal{S}\|_{\text{sum}} + \frac{\lambda_x}{K}\sum_{r=1}^{R}\sum_{k=1}^{K}\|a_r^{(k)}\|^K. \tag{26}$$

The following proposition shows that this nonconvex program is equivalent to (24) as long as the induced rank-bound is non-restrictive at the solution.

**Proposition 9.** *Suppose* $\{a_r^{*(1)}, \cdots, a_r^{*(K)}\}_{r=1,\ldots,R}, S^*$ *are optimal for the nonconvex program* (26)*, and define* $\mathcal{X}^* \stackrel{def}{=} \sum_{r=1}^{R}(a_r^{*(1)} \otimes \cdots \otimes a_r^{*(K)})$*. Then the point* $(\mathcal{X}^*, S^*)$ *is optimal for the problem* (24)*. Conversely, if* $(\mathcal{X}^*, \mathcal{S}^*)$ *is the minimizer of* (24)*, then the terms* $\{a_r^{*(1)}\}, \cdots, \{a_r^{*(K)}\}$ *from an atomic decomposition of* $\mathcal{X}^*$ *are optimal for* (26)*.*

The authors of [9] and [42] prove results similar to Proposition 9 in the case $K = 3$, and we include a proof in Appendix D for completeness.

The factorized regularizer in (26) can be viewed as a higher-order generalization of Burer-Monteiro factorization, which is popular in low-rank matrix recovery [13, 16, 40, 45]. Notice for $K = 2$, this term reduces to $\frac{\lambda_x}{2}(\|A\|_F^2 + \|B\|_F^2)$, which is equivalent to standard nuclear-norm regularization (as shown in (23)). The next section discusses how we can efficiently find a stationary point of (26).

## 5.2 First-Order Solvers

One of the drawbacks of (26) is that the non-smooth $\ell_1$-regularizer $\|S\|_{\text{sum}}$ prevents the use of pure gradient-based solvers. One can use proximal gradient methods but these can still be slow. However, as suggested in [15], the structure of the program allows us to smooth the problem through marginalization. Define

$$\varphi : \mathcal{X} \mapsto \min_S \frac{1}{2}\|\mathcal{X} + S - \mathcal{Z}\|_F^2 + \lambda_s\|S\|_{\text{sum}}. \tag{27}$$

Because we are using the least-squares loss and $\ell_1$-regularization, $\varphi$ is the (shifted) Moreau envelope of the $\ell_1$-norm, also known as the Huber loss [7]. The objective in (27) is strongly convex, so the minimum exists and is unique. In fact, it can be written in closed-form using the shrinkage operator:

$$(\text{shrink}(Y, \lambda))_{i_1,\cdots,i_K} \overset{\text{def}}{=} \text{sign}(Y_{i_1,\cdots,i_K}) \lfloor\, |Y_{i_1,\cdots,i_K}| - \lambda_s| \,\rfloor_+,$$

(where $\lfloor a \rfloor_+$ denotes the non-negative part of $a$), so that

$$\arg\min_S \frac{1}{2}\|\mathcal{X} + S - \mathcal{Z}\|_F^2 + \lambda_s \|S\|_{\text{sum}} = \text{shrink}(\mathcal{Z} - \mathcal{X}, \lambda_s).$$

Incorporating $\varphi$ into the convex program (6) preserves convexity and introduces differentiability. Combining $\varphi$ with the nonconvex program (26) yields a tractable, Lipschitz-differentiable problem that is amenable to first-order solvers. Formally, letting $\mathbf{a} = (a_r^{(k)})_{r=1,\ldots,R}^{k=1,\ldots,K}$, we solve the program

$$\min_{\mathbf{a}} \quad \frac{\lambda_x}{K} \sum_{r=1}^{R} \sum_{k=1}^{K} \|a_r^{(k)}\|^K + \varphi\left(\mathcal{A}(\mathbf{a})\right), \tag{28}$$

where

$$\mathcal{A} : \mathbf{a} \mapsto \sum_{r=1}^{R} (a_r^{(1)} \otimes \cdots \otimes a_r^{(K)})$$

Let $f(\mathbf{a})$ be the objective in (28), and let $S^{\mathbf{a}}$ be the minimizer in (27) (which implicitly depends on $\mathbf{a}$). Then $f$ is differentiable with gradient given by

$$\begin{aligned} \nabla_{a_r^{(k)}} f &= (\lambda_x \|a_r^{(k)}\|^{K-2}) a_r^{(k)} + \nabla_{a_r^{(k)}} \psi(\mathbf{a}) \Big|_{S^{\mathbf{a}}} \\ &= (\lambda_x \|a_r^{(k)}\|^{K-2}) a_r^{(k)} + \left(a_r^{(k)} C^T + S^{\mathbf{a}} - \mathcal{Z}\right) C, \end{aligned}$$

where $C = \left(A^{(K)} \odot \cdots \odot A^{(k+1)} \odot A^{(k-1)} \odot \cdots \odot A^{(1)}\right)$. We direct the reader to [34, Thm. 10.58] for proof of the first equality and to [1] for a derivation of the second.

## 5.3 Global Optimality of Certain Local Minima

Although it is possible to find a stationary point of (26) efficiently, it is not obvious that this stationary point approximates the global optimizer when $K \geq 3$. This contrasts with the case $K = 2$, where several works have shown that the local optima of (26) or similar models are all globally optimal, and non-optimal stationary points are avoidable.

The case $K \geq 3$ has received considerably less attention. However, we can use a recent result of [22] to show that certain local minima of (26) are globally optimal. The following proposition is a special case of [22, Thm. 15].

**Proposition 10** ([22]). *Let $\ell(\mathcal{X}, \mathcal{S})$ be once differentiable and jointly convex in $\mathcal{X}$ and $\mathcal{S}$, and let $R(\mathcal{S})$ be convex but not necessarily differentiable. Any local minimizer of the optimization problem*

$$\min_{\{a_r^{(1)}\},\cdots,\{a_r^{(K)}\},\mathcal{S}} \quad \ell\left(\sum_{r=1}^{R} (a_r^{(1)} \otimes \cdots \otimes a_r^{(K)}), \mathcal{S}\right) + R(\mathcal{S}) + \sum_{r=1}^{R}\sum_{k=1}^{K} \|a_r^{(k)}\|^K.$$

*such that $a_{r_0}^{(k_0)} = 0$ for some $r \in \{1, \cdots, R\}$ and $k \in \{1, \cdots, K\}$ is a global minimizer.*

Clearly, Proposition 10 applies to the model in (26). The condition that one of the factors $a_{r_0}^{k_0}$ is equal to zero is a natural condition for global optimality because it implies that the rank-bound the factorization induces is not too strict. Proposition 10 suggests that if $R$ is large enough, then the stationary points of (26) are good approximations of the global optimizers. This idea is discussed formally in [22], and we direct the interested reader there for a discussion of descent approaches that rigorously converge to global optimizers.

# 6 Tensor RPCA for Topic Modeling

In [4], the authors demonstrate that training many latent variable models can be reduced to a tensor decomposition problem. Current methods for obtaining this decomposition are generally based on the Tensor Power Method described in [4], and they all require the tensor being decomposed to have orthogonal factor matrices [46, 18, 4]. To meet this requirement, the underlying distributions must be linearly independent, and a numerically unstable whitening procedure must be implemented before the tensor decomposition takes place.

Tensor RPCA avoids these problems because it does not assume the factor matrices are orthogonal. Furthermore, it is also robust to sparsely distributed errors in the sample moment-tensor, which often arise from systematic errors in the sample moments [5]. In the following subsections, we outline how tensor RPCA can be used for topic modeling, specifically, for parameter estimation in Latent Dirichlet Allocation (LDA). Central to this approach is how Theorem 1 interpreted in this context provides provable performance guarantees, stating that as long as the number of errors in the sample moments is small, and the number of topics is small with respect to the vocabulary size, then tensor RPCA perfectly recovers the topic distributions with high probability.

## 6.1 Problem Setting and Existing Approaches

In the LDA model, every document is a mixture of topics, and these mixtures follow a Dirichlet distribution $\mathrm{Dir}(\beta)$. Here, $\beta \in \mathbb{R}_{++}^K$ is a vector of $K$ model parameters, where $K$ is the number of topics. The probability density of this distribution over the simplex $\Delta^{K-1}$ is given by [10]

$$p_\beta(h) = \frac{\Gamma(\beta_0)}{\prod_{i=1}^K \Gamma(\beta_i)} \prod_{i=1}^K h_i^{\beta_i - 1}, \quad h \in \Delta^{K-1}, \quad \beta_0 := \sum_{i=1}^K \beta_i.$$

To form a document under this model, we first draw a topic mixture $h = (h_1, \cdots, h_K) \sim \mathrm{Dir}(\beta)$, and conditioned on this mixture, we independently draw $\ell$ words $w_1, \cdots, w_\ell$ from the distribution $\sum_{i=1}^K h_i \nu_i$, where $\nu_i \in \Delta^{d-1}$ represents the distribution over the vocabulary corresponding to the $i^{th}$ topic. The words $w_i$ follow a one-hot vector encoding, so $w_i = e_i$ (where $e_i$ is a canonical basis vector) if and only if the $i^{th}$ word of the document is $w_i$. The following theorem relates low-order moments of this model to the topic distributions:

**Theorem 11** ([3, 4]). *Let*

$$M_1 \overset{def}{=} \mathbb{E}[w_1], \qquad M_2 \overset{def}{=} \mathbb{E}[w_1 \otimes w_2] - \frac{\beta_0}{\beta_0 + 1} M_1 \otimes M_1,$$

$$\mathcal{M}_3 \overset{def}{=} \mathbb{E}[w_1 \otimes w_2 \otimes w_3] - \frac{\beta_0}{\beta_0 + 2}(\mathbb{E}[w_1 \otimes w_2 \otimes M_1] + \mathbb{E}[w_1 \otimes M_1 \otimes w_2] \\ + \mathbb{E}[M_1 \otimes w_1 \otimes w_2]) + \frac{2\beta_0^2}{(\beta_0+2)(\beta_0+1)} M_1 \otimes M_1 \otimes M_1.$$

*Then*

$$M_2 = \sum_{i=1}^{K} \frac{\beta_i}{(\beta_0+1)\beta_0} \nu_i \otimes \nu_i, \qquad \mathcal{M}_3 = \sum_{i=1}^{K} \frac{2\beta_i}{(\beta_0+2)(\beta_0+1)\beta_0} \nu_i \otimes \nu_i \otimes \nu_i \tag{29}$$

Hence, the topic distributions $\nu_i$ can be discovered from the rank-one decompositions of $M_2$ and $\mathcal{M}_3$; decomposition of $M_2$ alone is not sufficient since there is not a *unique* rank-$K$ decomposition of a symmetric matrix, whereas the tensor CP decomposition is often unique, cf. Thm. 13. As described in [4], this type of structure exists in the low-order moments of numerous latent variable models. The results presented in this paper extend to these problems as well.

## 6.2 Theorem 1 for Provable Topic Modeling

In the context of topic modeling, we would like to recover the population moment-tensor $\mathcal{M}_3$ in (29) and its rank-one decomposition even though the empirical moment-tensor has entries with large errors. Theorem 1 shows that exactly recovering the population moment-tensor is possible when the number of topics and the number of errors are bounded.

**Corollary 12.** *Let $\mathcal{Z} \in \mathbb{R}^{d\times d\times d}$ be the empirical third-order moment-tensor. Suppose the population moment-tensor $\mathcal{M}_3 \in \mathbb{R}^{d\times d\times d}$ satisfies $\mu(\mathcal{M}_3) \le \mu_0$, and that the true number of topics is* $\operatorname{rank}_{\mathrm{CP}}(\mathcal{M}_3) = K$*. Let tensor $\mathcal{S} \in \mathbb{R}^{d\times d\times d}$ be the tensor of discrepancies*[1] *between the empirical and population moment-tensors. Suppose $\mathcal{S}$ has a support set $\Omega$ that is uniformly distributed among all sets of cardinality $m$, and let $n = d^3 - m$. Then there exists a positive constant $c$ so that tensor RPCA with $\lambda = (3d)^{-1/2}$ exactly recovers $\mathcal{M}_3$ and $\mathcal{S}$ with probability $1 - d^{-1-c}$, provided that*

$$K \le \rho_r \left( \frac{n}{3d\log(n)\alpha_0^4\mu_0^2} \right)^{1/2} \quad \textit{and} \quad m \le \rho_s d^3.$$

*Proof.* Recall that $\operatorname{rank}_{\mathrm{CP}}(\mathcal{M}_3) \in [\overline{r}(\mathcal{M}_3), \overline{r}(\mathcal{M}_3)^2]$, so our assumption implies $\overline{r}(\mathcal{M}_3) \le \rho_r \left( \frac{n}{3d\log(n)\alpha_0^4\mu_0^2} \right)^{1/2}$. Applying Theome 1 proves the result. □

Hence, if the number of topics is much less than the size of our vocabulary, $d$, then the actual third-order moment can be exactly recovered.

Of course, this result says nothing about the recovery of rank-one components of the moment-tensor $\mathcal{M}_3$, which are what reveal the topic distributions. Also, it is impractical to work with tensors of size $d \times d \times d$, as the vocabulary size is generally extremely large. We address each of these problems in the following sections.

[1] In practice, a robust version of RPCA, much like the Lagrangian formulation used in (24), can be used, which distinguishes ubiquitious small-magnitude discrepancies from rare but large discrepancies. There are matrix RPCA results for this case which cannot promise exact recovery, but do guarantee recovery up to the level of the small-magnitude noise.

## 6.3 Identifiability

By solving the nonconvex formulation of tensor RPCA (26), we implicitly solve for a rank-one decomposition of the moment-tensor. However, if this CPD is not unique, it is unclear whether the recovered rank-one factors correspond to the topic distributions. Fortunately, a tensor's CPD is unique under mild conditions.

**Theorem 13.** *[27] Let $k_\nu$ be the maximum value such that the vectors in any subset of $\{\nu_r\}_{r=1}^K$ of size $k_\nu$ are linearly independent. If $3k_\nu \geq 2K+2$, then the CPD of $\mathcal{M}_3 = \sum_{r=1}^R \nu_r \otimes \nu_r \otimes \nu_r$ is unique up to permutation and scaling of the topic distributions $\nu_r$.*

The condition in Theorem 13 is weaker than requiring the distributions to be linearly independent, which is necessary for most existing approaches [3, 4]. Further information on using Kruskal's theorem to establish identifiability results in latent structure models can be found in [2].

For any CPD, it is possible to form an equivalent CPD by permuting and rescaling the columns of the factor matrices. Precisely, if some tensor $\mathcal{X} = \sum_{r=1}^R a_r \otimes b_r \otimes c_r$, then $\mathcal{X} = \sum_{r=1}^R k_1 a_{\pi(r)} \otimes k_2 b_{\pi(r)} \otimes k_3 c_{\pi(r)}$ as well, where $\pi$ is any permutation on the set $\{1, 2, \cdots R\}$ and $k_1 k_2 k_3 = 1$. However, neither of these operations affect the topic distributions found from $\mathcal{M}_3$ because we constrain each rank-one factor of $\mathcal{M}_3$ to be symmetric (i.e., of the form $\nu_r \otimes \nu_r \otimes \nu_r$). This restriction disallows rescaling. Permuting the topic distributions does not change the distributions, so equivalence up to scaling and permutation still ensures that the topic distributions are well-defined.

## 6.4 Dimensionality Reduction, Whitening, and Oversampling

It is often not feasible to decompose tensors of size $d \times d \times d$ when the vocabulary is large. To overcome this, existing works apply a dimensionality reduction technique so that the topic distributions can be recovered from the rank-one factors of a smaller tensor [3, 4]. This dimensionality reduction step is closely related to the whitening procedure presented in [3, 4] that is used to orthogonalize the factor matrices of a tensor so that orthogonal-decomposition algorithms, such as the Tensor Power Method, are applicable. In this section, we outline the whitening procedure, discuss its numerical instability, and show how a similar but stable procedure can be used for dimensionality reduction in tensor RPCA.

Suppose the topic distributions we seek are linearly independent. Let $M_2$ be the empirical second-order moment in (29), and let $M_2 =: U\Sigma V^T$ be its (skinny) singular value decomposition. Define the whitening matrix as $W \stackrel{\text{def}}{=} \Sigma^\dagger U^T \in \mathbb{R}^{K \times d}$, where $\Sigma^\dagger$ is the pseudo-inverse of $\Sigma$. The whitened third-order empirical moment is then [3, 4]

$$\widetilde{\mathcal{M}}_3 \stackrel{\text{def}}{=} (W, W, W) \cdot \mathcal{M}_3 = \sum_{r=1}^K \lambda_r (W\nu_i \otimes W\nu_i \otimes W\nu_i) \in \mathbb{R}^{K \times K \times K}$$

The components $W\nu_i$ are orthogonal and can be found by decomposing $\widetilde{\mathcal{M}}_3 \in \mathbb{R}^{K \times K \times K}$, which is much smaller than the original tensor $\mathcal{M}_3 \in \mathbb{R}^{d \times d \times d}$. The factors of $\mathcal{M}_3$ can then be found by applying the inverse whitening transform, $\mathcal{M}_3 = (W^\dagger, W^\dagger, W^\dagger) \cdot \widetilde{\mathcal{M}}_3$, where $W^\dagger = US$. Multiplying by the pseudo-inverse $W^\dagger$ is numerically unstable if $W$ has a large condition number, which is common for real-world data sets, and this introduces unnecessary error into the computation.

For tensor RPCA, we can perform the same dimensionality reduction without whitening. Tensor RPCA does not require the tensor to be orthogonally decomposable, so we can drop the assumption that the topic distributions are linearly independent. The transformation $Q \overset{\text{def}}{=} I_{K\times d}U^T$ performs the same dimensionality reduction as the whitening transform: $\widehat{\mathcal{M}}_3 \overset{\text{def}}{=} (Q,Q,Q)\cdot\mathcal{M}_3 \in \mathbb{R}^{K\times K\times K}$. However, $Q$ does not orthogonalize the factor matrices of $\mathcal{M}_3$ as this is not required, and $Q$ is perfectly conditioned with condition number 1. The factors of $\mathcal{M}_3$ can be recovered from the factors of $\widehat{\mathcal{M}}_3$ by using the inverse transformation matrix $Q^\dagger = Q^T I_{d\times K}$, and because $Q$ is perfectly conditioned, this inverse transformation does not introduce any errors due to numerical instability.

Instead of reducing the dimension to equal the number of topics, RPCA performs better with oversampling. The results of Corollary 12 require the tensor we decompose to have sufficiently large dimensions compared to the number of expected topics, so we can use the transformation matrix $Q' \overset{\text{def}}{=} I_{K'\times d}U^T$, where $K'$ is chosen so that $K \leq \rho_r\left(\frac{K'^3-m}{3K'\log(K'^3-m)\alpha_0^4\mu_0^2}\right)^{1/2}$, in accordance with Corollary 12.

# 7 Numerical Experiments

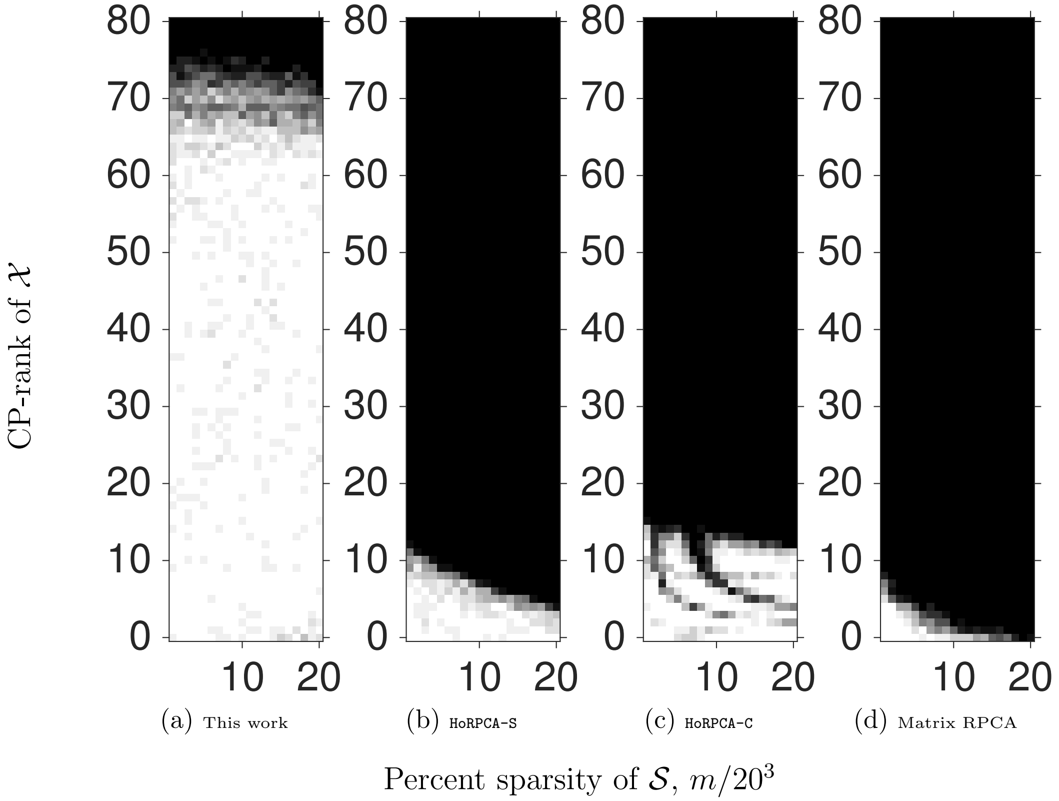
We compare our model to existing methods for tensor and matrix RPCA on synthetic data and the escalator video dataset of [30]. Our experiments demonstrate that our model significantly outperforms existing methods for tensor and matrix RPCA. We also see that we perform much better than the guarantees given in Theorem 1. Most remarkably, our model is able to recover tensors whose rank is much larger than its side lengths. In this regime, the Tucker-rank of the tensor is no longer an appropriate measure of the complexity of the data, and many existing methods for tensor RPCA [19, 24] are ineffective.

## 7.1 Experiments on Synthetic Data

For each trial, we create an order-three, cubic dataset that can be represented as the sum of a low-rank tensor and a sparse tensor. To form the low-rank component, we randomly generate three factor matrices $A, B, C \in \mathbb{R}^{20\times R}$, with each entry drawn i.i.d. $\mathcal{N}(0,1)$. We then form the low-rank component as $\mathcal{X} = \sum_{r=1}^{R}(a_r \otimes b_r \otimes c_r)$. We set the rank bound to be $R+10$ in our algorithm.

To form the sparse component, we make the tensor $\mathcal{S} \in \mathbb{R}^{20\times 20\times 20}$ with support chosen uniformly at random without replacement such that there are $m$ nonzeros, and non-zero entries drawn i.i.d. $\mathcal{N}(0,1)$. Our "observed" dataset is then $\mathcal{Z} = \mathcal{X} + \mathcal{S}$. We vary both the rank $R$ and sparsity $m$.

For each test, we perform 16 trials and measure the error between the recovered low-rank component $\mathcal{X}'$ and the actual low-rank component $\mathcal{X}$ using the relative least-squares loss $\frac{\|\mathcal{X}'-\mathcal{X}\|_F}{\|\mathcal{X}\|_F}$, declaring "exact recovery" when this error is below $10^{-3}$. We fit our model using L-BFGS as implemented in [35], maintaining 10 iterations in memory, and we stop each trial after 1,000 iterations. For our parameters, we set $\lambda_x = 10^{-5}$ and $\lambda_s = 10^{-3}$. Both are small



(a) This work (b) HoRPCA-S (c) HoRPCA-C (d) Matrix RPCA

Figure 1: A comparison of RPCA methods for recovering the decomposition $\mathcal{Z} = \mathcal{X} + \mathcal{S}$ with $\mathcal{Z} \in \mathbb{R}^{20\times 20\times 20}$. A pixel is colored white if $\mathcal{X}$ is recovered exactly. Each pixel represents the average of 16 trials. Algorithms (b), (c), and (d) are ill-posed when the CP-rank is greater than 20.

because we do not expect there to be any noise in $\mathcal{Z}$. Our results are shown in Figure 1, along with the results of the same experiment using matrix RPCA and two existing tensor RPCA methods.

The ranks reported in the figure are upper bounds, as it is possible for a certain $\mathcal{X} = \sum_{r=1}^{R}(a_r \otimes b_r \otimes c_r)$ to admit a lower-rank CP-decomposition. The Tucker-rank of the low-rank component is $(R, R, R)$ for $R \leq 20$, and it is $(20, 20, 20)$ for $R \geq 20$. This is explicitly checked for each trial.

The two other tensor-based models, `HoRPCA-C` [19] and `HoRPCA-S` [19, 24], use the sum-of-nuclear-norms (SNN) regularizer, and `HoRPCA-S` is one of the only provable method for recovering low-rank tensors outside of this work [24]. `HoRPCA-S` solves the problem

$$\min_{\mathcal{X},\mathcal{S}} \quad 20\sum_{i=1}^{K} \|\mathcal{X}'_{(i)}\|_* + \|\mathcal{S}'\|_{\text{sum}}, \quad \text{subject to:} \quad \mathcal{X}' + \mathcal{S}' = \mathcal{Z},$$

See (9) for their recovery guarantees. In our experiments, we see that `HoRPCA-S` performs worse than our tensor RPCA, and marginally better than matrix RPCA. When $R \geq 20$, `HoRPCA-S` is an ill-posed problem, because each matricization of $\mathcal{X}$ has full rank.

HoRPCA-C exhibits similar behavior. This program is defined as

$$\min_{\mathcal{X}',\mathcal{S}'} \quad \|\mathcal{S}'\|_{\text{sum}} \quad \text{subject to:} \quad \mathcal{X}' + \mathcal{S}' = \mathcal{Z}, \quad \text{rank}(\mathcal{X}'_{(i)}) \leq r_i.$$

Although nonconvex, it has been shown to outperform other methods for tensor RPCA, including HoRPCA-S [19]. However, this model still suffers from the effects of tensor matricization. For $R \geq 20$ in our experiments, HoRPCA-C is an ill-posed problem. For our implementation of HoRPCA-C, we set $r_i = R + 1$ for all $i$. These rank bounds are much tighter than the rank bound $(R+10)$ we used for our model, but because $r_i$ cannot be larger than 20, it would not make sense to use the rank bound $R + 10$ for the components of the Tucker-rank.

For the matrix RPCA in our experiments, we matricize the tensors and use the variational approach to RPCA developed in [6].[2] The program we solve is

$$\min_{\mathcal{X}',\mathcal{S}'} \quad \max(\|\mathcal{X}'_{(1)}\|_*, \lambda\|\mathcal{S}'_{(1)}\|_{\text{sum}}) \quad \text{s.t.} \quad \frac{1}{2}\|\mathcal{X}'_{(1)} + \mathcal{S}'_{(1)} - \mathcal{Z}_{(1)}\|_F^2 \leq \epsilon.$$

One of the benefits of this approach is that we can choose $\lambda$ optimally because we know the matrices we would like to recover. We set $\lambda = \frac{\|\mathcal{X}_{(1)}\|_*}{\|\mathcal{S}_{(1)}\|_1}$, and because we do not expect there to be any noise in $\mathcal{Z}$, we choose $\epsilon = 10^{-5}$.

## 7.2 Tensor RPCA for Background Subtraction

One of the most natural applications for RPCA is in background subtraction. In this section, we use our model to identify subjects in the "escalator-video" dataset provided by [30]. This dataset is challenging for background-subtraction models because it contains three moving parts: a time-stamp, escalators, and the subjects. A strong model would be able to recognize that the motion of the escalators and the time-stamp is periodic, so these features belong to the low-rank component of the dataset, and the unpredictable motion of the subjects should be extracted into the sparse component.

In Figure 2, we compare the performance of our tensor RPCA model and matrix RPCA for identifying the background in surveillance video. The video consists of 200 frames of size $130 \times 160$, which we store in a tensor of size $130 \times 160 \times 200$ or a matrix of size $20800 \times 200$. For tensor RPCA, we set the parameters $\lambda_x = 30$, $\lambda_s = 0.1$, and the rank-bound $R = 50$. For matrix RPCA, we choose $\lambda = 0.02$ and $\epsilon = 9 \times 10^3$. All of these parameters were chosen after careful tuning. The parameter $\epsilon$ was chosen to match the noise-level in the tensor RPCA solution to make the results more comparable. We see in Figure 2 that tensor RPCA recovers a qualitatively superior decomposition, with the sparse component containing the subjects and very little of the stairs, and the low-rank component sufficiently "sharp." In contrast, the low-rank component found by matrix RPCA appears more smoothed and has "ghosts" where the subjects should be removed, while the sparse component contains a significant amount of the stairs.

Even more impressive is the quantitative difference between the decompositions. The numerical rank of the low-rank component found using tensor RPCA is 48, and the sparse

[2] Matrix RPCA code available at `https://github.com/stephenbeckr/fastRPCA/`

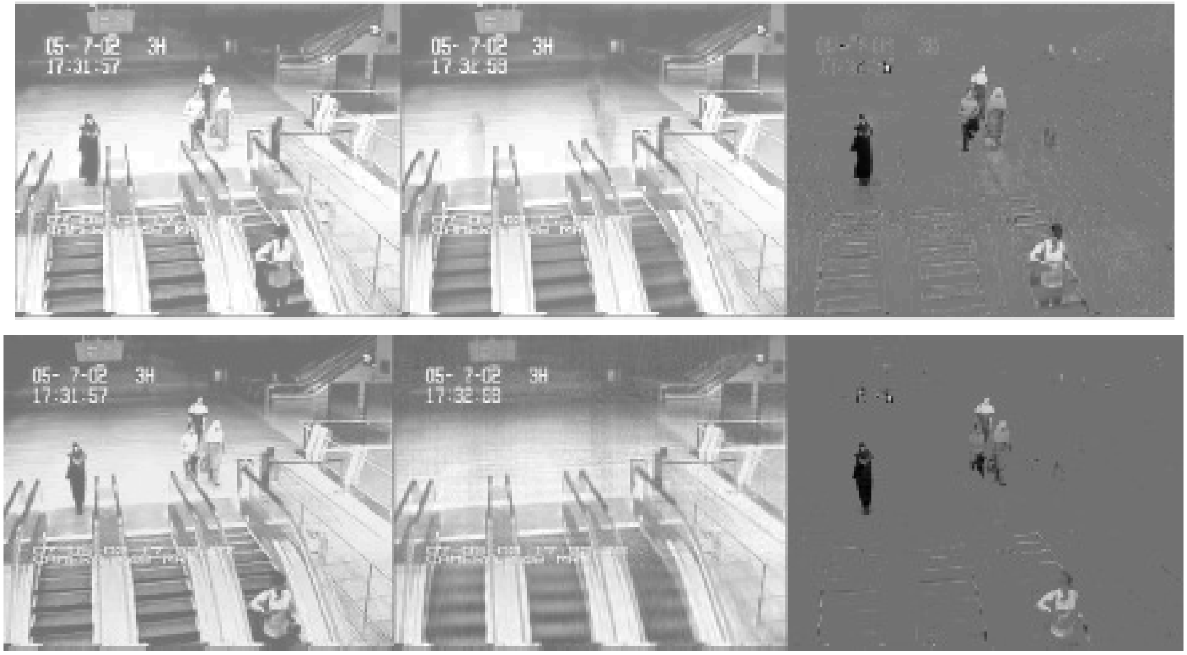


Figure 2: The results of using (top) matrix RPCA and (bottom) tensor RPCA for background subtraction. From left to right: original image, low-rank component ($\mathcal{X}$), and sparse component ($\mathcal{S}$). We see that tensor RPCA can more precisely extract the subjects into the sparse component and leave the moving escalator in the low-rank component. Frame 5 shown.

component is 5.5%-sparse. For matrix RPCA, the recovered low-rank component has rank 58, and the sparse component is 49.3%-sparse. It is clear that tensor-RPCA recovers a sparse component with significantly fewer non-zero entries than the sparse component that matrix RPCA recovers. To compare the low-rank components, it is better to compare degrees of freedom than the actual ranks. A tensor of size $d_1 \times d_2 \times d_3$ with CP-rank $r$ has $r(d_1 + d_2 + d_3)$ degrees of freedom, so tensor RPCA finds a low-rank component with 23,520 degrees of freedom, and matrix RPCA finds a low-rank component with 1,218,000 degrees of freedom, which is an enormous difference in the complexity of the solutions.

## 7.3 Topic Modeling Experiments

Using the `20newsgroups` dataset [28], we compare our tensor RPCA strategy to the tensor approach of [4] and the standard variational Bayes approach. We use a training set of 11,314 documents, and compare our results on a testing set of 7,532 documents. Our vocabulary consists of the 1,000 most common words in the training-set corpus, disregarding words that occur in more that 95% of the documents and those that occur in fewer than five documents. We also removed all English stopwords given by the University of Glasgow's stopword list. We look for five topics to discover, hoping to discover broad topics related to technology, recreation, science, politics, and religion. We do not include headers, footers, or quotes.

For the variational Bayes tests, we use SciKit-Learn's implementation of Latent Dirichlet Allocation, with hyperparameters $\alpha = 0.1, \beta = 0.1, \kappa = 0.7$, and $\tau_0 = 50$. We run the solver for twenty iterations. We use code provided by the authors of [3] for our comparison to their

|  | Perplexity | Average Observed Coherence |
|---|---|---|
| Variational Bayes | 1417 | 0.033 |
| Tensor CPD | 1503 | 0.006 |
| Tensor RPCA | 1309 | 0.007 |

Figure 3: A comparison of topic models on the 20newsgroups dataset. We use SciKit-Learn's implementation of variation Bayes for LDA and the method of [3] for the tensor CPD. The observed coherence for each topic is calculated using the method of [29], and we present the average observed coherence over all five topics.

method. For these tests, we set $\alpha_0 = 0.01$ and the number of topics $K = 5$. For tensor RPCA, we oversample, setting $K' = 10$, and we choose the other parameters to $\lambda = 10^{-3}$ and $\mu = 10^{-8}$. The topics that each approach finds are included in Appendix A.

The decomposition from tensor RPCA has numerical rank five, showing that it discovers five topics in the data. It finds 87 sparse errors. We quantitatively compared our results using the perplexity on the test set, defined for each word $w$ in the test set as

$$\text{perplexity}(w) = \exp\left(-\frac{\mathcal{L}(w)}{d}\right),$$

where $d$ is the size of the vocabulary and $\mathcal{L}(w) = \log p(w|\Phi, \alpha)$ is the log-likelihood of the test set given the estimated topic-matrix $\Phi$ and hyperparameter $\alpha$. To quantitatively measure the interpretability of the topics, we used the observed coherence measure from [29]. A comparison using these measures is shown in Figure 3. Qualitatively, it appears that variational Bayes has the best topic coherence, but it also finds meaningless topics, grouping low-frequency words or numbers together, which are not article topics. The tensor-based approaches do not exhibit this problem.

The tensor decomposition step was a bit faster using the approach of [4], taking 3.46 seconds for the decomposition step to converge compared to tensor RPCA's 5.13 seconds. We measure convergence with respect to the relative change in the low-rank component, $\frac{\|\mathcal{X}_k - \mathcal{X}_{k-1}\|_F}{\|\mathcal{X}_k\|_F}$. We set the tolerance to $10^{-4}$ for both approaches. Both tensor-based methods are significantly faster than variational Bayes, which took 44.71 seconds to perform twenty iterations.

# 8 Conclusion

Our guarantees show that tensor RPCA with atomic-norm regularization outperforms matrix-based RPCA and RPCA algorithms based on matricization. Although the atomic norm is generally intractable, our use of a higher-order generalization of Burer-Monteiro factorization allows us to derive a nonconvex program equivalent to tensor RPCA. Our nonconvex model can be fit efficiently using any first-order optimizer. While convergence to a global optimum is not generally ensured, we provide sufficient conditions for a local minimum to be globally optimal which can be verified *ex post facto*.

As an algorithm for the low-rank decomposition of tensors, tensor RPCA can be used for training many latent variable models, including LDA. In this context, our main result offers performance guarantees for estimating relevant parameters. Tensor RPCA offers many improvements over existing methods for decomposing moment-tensors. Notably, it does not require the underlying distributions to be linearly independent, and it avoids the unstable whitening step required by many existing methods. Tensor RPCA also provably recovers the population moment-tensor in the presence of sparsely distributed errors in the sample moments.

Empirically, our tensor RPCA significantly outperforms matrix RPCA as well as existing implementations of RPCA that use sum-of-nuclear-norm regularization. Our approach to tensor RPCA is also able to recover tensors whose CP-rank greatly exceeds all of its side lengths, a regime where sum-of-nuclear-norm models are ill-posed. Our results suggest that analyzing low-rank tensor recovery in terms of the Tucker-rank does not yield tight performance bounds, and future work might investigate performance guarantees in terms of the CP- or atomic-rank.

# References

[1] E. Acar, D. M. Dunlavy, and T. G. Kolda. A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics*, 25(2):67–86, 2011.

[2] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6), 2009.

[3] A. Anandkumar, D. Foster, D. Hsu, S. Kakade, and Y. Liu. Two SVDs suffice: Spectral decompositions for probabilistic topic modeling and latent Dirichlet allocation. In *NIPS*, 2012.

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *JMLR*, pages 2773–2832, 2014.

[5] A. Anandkumar, P. Jain, Y. Shi, and U. N. Niranjan. Tensor vs matrix methods: Robust tensor decomposition under block sparse perturbations. *arXiv preprint arXiv:1510.04747*, 2015.

[6] A. Aravkin, S. Becker, V. Cevher, and P. Olsen. A variational approach to stable principal component pursuit. In *Uncertainty in Artificial Intelligence*, Quebec City, 2014.

[7] A. Aravkin, J. Burke, D. Drusvyatskiy, M. Friedlander, and S. Roy. Level-set methods for convex optimization. *arXiv preprint*, 2016.

[8] B. Barak and A. Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *JMLR Workshop and Conference Proceedings*, volume 49, pages 417–445, 2016.

[9] J. Bazerque, G. Mateos, and G. Giannakis. Rank regularization and bayesian inference for tensor completion and extrapolation. *IEEE Signal Processing*, 2013.

[10] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

[11] H. Botao, A. Zhang, and G. Cheng. Sparse and low-rank tensor estimation via cubic sketchings. *arXiv preprint arXiv:1801.09326*, 2018.

[12] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58(1):1–37, 2011.

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

[14] P. Comon and C. Jutten. Handbook of blind source separation: Independent component analysis and applications. *Elsevier*, 2010.

[15] D. Driggs, S. Becker, and A. Aravkin. Adapting regularized low-rank recovery models for parallel architectures. *SIAM J. Sci. Comp.*, 41(1), 2019.

[16] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE, 2001.

[17] S. Friedland and L.-H. Lim. Nuclear norm of higher-order tensors. *Mathematics of Computation*, 2016.

[18] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Proceedings of the 28th Conference on Learning Theory*, pages 797–842, 2015.

[19] D. Goldfarb and Z. Qin. Robust low-rank tensor recovery: Models and algorithms. *SIAM Journal on Matrix Analysis and Applications*, 2017.

[20] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theory*, 57(3):1548?–1566, 2011.

[21] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Physical Review Letters*, 105(15), 2010.

[22] B. D. Haeffele and R. Vidal. Global optimality in tensor factorization, deep learning, and beyond. *ArXiv:1506.07540*, 2015.

[23] C. Hillar and L.-H. Lim. Most tensor problems are np-hard. *J. ACM*, 6, 2013.

[24] B. Huang, C. Mu, D. Goldfarb, and J. Wright. Provable low-rank tensor recovery. *Optimization-Online*, page 4252, 2014.

[25] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint*, 2015.

[26] T. Kolda and B. Bader. Tensor decompositions and application. *SIAM Review*, pages 455–500, 2009.

[27] J. Kruskal. Rank, decomposition, and uniqueness for 3-way and N-way arrays. In *Multiway Data Analysis*, pages 7–18, 1989.

[28] K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

[29] J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530—-539, 2014.

[30] L. Li, W. Huang, I. Gu, and Q. Trian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transaction on Image Processing*, 2004.

[31] C. Lu, Y. Chen, J. Feng, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[32] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *arXiv:1804.03728*, 2018.

[33] A. Potechin and D. Steurer. Exact tensor completion with sum-of-squares. *arXiv:1702.06237*, 2017.

[34] R. Rockafellar and R. Wets. *Variational Analysis.* Springer Berlin Heidelberg, 2009.

[35] M. Schmidt. minFunc: unconstrained differentiable multivariate optimization in Matlab. 2005.

[36] P. Shah, N. Rao, and G. Tang. Sparse and low-rank tensor decomposition. In *Proceedings of the 28th Advance in Neural Information Processing Systems*, pages 2548–2556, 2015.

[37] N. Sidiropoulos, L. D. Lathauwer, X. Fu, K. Huang, E. Papalexakis, and C. Faloutsos. Tensor decompositions for signal processing and machine learning. *IEEE Trans. Sig. Proc.,* to appear, 2017.

[38] V. D. Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J. Matrix Anal. Appl.*, 30:1084–1127, 2008.

[39] A. Sobral, S. Javed, S. K. Jung, T. Bouwmans, and E. Zahzah. Online stochastic tensor decomposition for background subtraction in multispectral video sequences. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 106–113, 2015.

[40] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. In *NIPS*, 2005.

[41] W. Sun and L. Li. STORE: Sparse tensor response regression and neuroimaging analysis. *Journal of Machine Learning Research*, 2017.

[42] G. Tang and P. Shah. Guaranteed tensor decomposition: A moment approach. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1491–1500, 2015.

[43] R. Tomioka and T. Suzuki. Spectral norm of random tensors. *arXiv preprint*, 2014.

[44] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 1966.

[45] M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2015.

[46] Y. Wang, H.-Y. Tung, A. Smola, and A. Anandkumar. Fast and guaranteed tensor decomposition via sketching. In *NIPS*, 2015.

[47] M. Yuan and C.-H. Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16:1031–1068, 2016.

# Appendices

## A Topic Model Results

| Topic | Associated words |
|---|---|
| 1 | 10 00 space 20 new 15 25 12 11 14 gun 30 000 16 1993 17 18 year 13 president april 50 mr national states |
| 2 | edu file use windows com program thanks drive like available software mail does using files version data know card problem window ftp email info used |
| 3 | key use government chip public used encryption keys like new law security clipper information bike bit privacy using number data don private does technology make |
| 4 | people don just think like know time god good say did said does way right ve believe make going really things want didn ll years |
| 5 | ax max b8f g9v a86 pl 145 1d9 0t 34u 1t 3t giz bhj wm 2di 75u 2tm cx bxn 7ey w7 chz sl 0d |

Table 1: Topics found using variational Bayes for LDA parameter estimation

| Topic | Associated words |
|---|---|
| 1 | effect sorry articles come unix current main trying dos doesn involved encryption likely report day news sense administration files interesting lots mode suggest sources key |
| 2 | dos perfect key thinking goes knowledge likely doesn saying didn rights ways best sale times reason think jewish got makes wanted good switch die things |
| 3 | likely news key driver times got used version thinking knowledge dos gov theory makes problems useful years effect san needed ways local 100 card windows |
| 4 | knowledge doesn theory dos likely key wire apparently main used needed age hi information high looks administration cards approach function programs input wanted meaning encryption |
| 5 | b8f windows used maybe card version wire programs come israeli graphics best dos rights age problem articles difficult manual keyboard distribution usually happen little switch |

Table 2: Topics found using the tensor-based method of moments from [3]. ALS was used for the tensor decomposition.

| Topic | Associated words |
|---|---|
| 1 | b8f news 000 card likely years games got teams times yes san dos reason 100 subject key gas thinking second didn agree things ways rights |
| 2 | likely key knowledge dos effect doesn thinking perfect used got theory news times version needed ways wanted rights reason makes useful didn main saying wire |
| 3 | effect sorry articles unix current encryption news san 000 involved conference come day got report suggest assume sense 100 1d9 department united main printer office |
| 4 | b8f theory doesn dos card maybe age driver games gas used files wire lord high function main source problems approach term looks programs letter apparently |
| 5 | theory effect knowledge doesn main wire age encryption high files looks hi information apparently administration function programs input cards final needed approach used solution sense |

Table 3: Topics found using the tensor-RPCA based method-of-moments in this work

# B Equivalence of Bernoulli and Uniform Sampling Models

We would like to show that if conditions (14)(d) and (e) hold with high probability when the support of $\mathcal{S}$ follows a Bernoulli distribution with parameter $2\rho$, then the same conditions hold with high probability when the support of $\mathcal{S}$ is uniformly distributed over all sets of cardinality $m$. We also need to show that the conclusions of Lemma 8 hold under the Bernoulli model as well as the uniform model. We first prove equivalence between the Bernoulli and uniform models for any event whose probability decreases with $|\Omega|$, and then we show that conditions (14)(d) and (e) and Lemma 8 satisfy this property.

**Lemma 14.** *Let $E$ be any event whose probability decreases as $|\Omega|$ increases. If $E$ holds with high probability when $\Omega$ follows the Bernoulli distribution of* (19)*, then $E$ holds with high probability when the support of $\Omega$ is uniformly distributed over all sets of cardinality $m$, and the converse holds as well.*

*Proof.* Our proof is similar to [12, App. 7]. Let $\mathbb{P}_{\mathrm{Ber}(\rho)}$ and $\mathbb{P}_{\mathrm{Unif}(k)}$ denote the probabilities calculated under the Bernoulli and uniform models, respectively. We begin by showing that if $E$ holds with high probability under the Bernoulli model, then it also holds with high probability under the uniform model.

$$\begin{aligned}\mathbb{P}_{\mathrm{Ber}(\rho)}\left(E\right) &= \sum_{k=0}^{d_1d_2d_3} \mathbb{P}_{\mathrm{Ber}(\rho)}\left(E\big||\Omega|=k\right)\mathbb{P}_{\mathrm{Ber}(\rho)}\left(|\Omega|=k\right)\\ &\leq \sum_{k=0}^{m-1}\mathbb{P}_{\mathrm{Ber}(\rho)}\left(|\Omega|=k\right) + \sum_{k=m}^{d_1d_2d_3}\mathbb{P}_{\mathrm{Unif}(k)}\left(E\right)\mathbb{P}_{\mathrm{Ber}(\rho)}\left(|\Omega|=k\right)\\ &\leq \mathbb{P}_{\mathrm{Ber}(\rho)}\left(|\Omega|<m\right) + \mathbb{P}_{\mathrm{Unif}(m)}(E).\end{aligned}$$

Here, we have used the facts that the distribution of $\Omega$ conditioned on its cardinality is uniform, and that $\mathbb{P}_{\mathrm{Unif}(k)}\left(E\right) \leq \mathbb{P}_{\mathrm{Unif}(m)}\left(E\right)$ for all $k \geq m$. This implies that

$$\mathbb{P}_{\mathrm{Unif}(m)}\left(E\right) \geq \mathbb{P}_{\mathrm{Ber}(\rho)}\left(E\right) - \mathbb{P}_{\mathrm{Ber}(\rho)}\left(|\Omega|<m\right).$$

Choosing $\rho = \frac{m}{d_1d_2d_3} + \epsilon$ for $\epsilon > 0$, we have $\mathbb{P}_{\mathrm{Ber}(\rho)}(|\Omega|<m) \leq e^{-\frac{\epsilon^2 d_1d_2d_3}{2\rho}}$. This proves that $E$ holds with high probability under the uniform model.

The converse holds as well. Suppose $E$ holds under the uniform model. Then

$$\begin{aligned}\mathbb{P}_{\mathrm{Ber}(\rho)}\left(E\right) &\geq \sum_{k=0}^{m}\mathbb{P}_{\mathrm{Ber}(\rho)}\left(E\big||\Omega|=k\right)\\ &\geq \mathbb{P}_{\mathrm{Unif}(m)}(E)\sum_{k=0}^{m}\mathbb{P}_{\mathrm{Ber}(\rho)}(|\Omega|=k)\\ &= \mathbb{P}_{\mathrm{Unif}(m)}(E)\mathbb{P}_{\mathrm{Ber}(\rho)}(|\Omega|\leq m).\end{aligned}$$

Choosing $m$ large enough ensures that $\mathbb{P}(|\Omega|>m)$ is exponentially small, so $E$ holds with high probability under the Bernoulli model as well. □

All that is left is to show that conditions (14)(d) and (e) and Lemma 8 satisfy the assumptions of the previous lemma.

**Lemma 15.** *Let $\mathcal{S}_0$ have support set $\Omega_0$, and let $\mathcal{S}_0'$ have support set $\Omega_0' \subsetneq \Omega_0$. Let $\mathcal{W}^{S_0} = \lambda\mathcal{P}_{\mathcal{X}^\perp}(\mathcal{P}_{\Omega_0} - \mathcal{P}_{\Omega_0}\mathcal{P}_{\mathcal{X}}\mathcal{P}_{\Omega_0})^{-1}\mathrm{sgn}(S_0)$ and $G_0 = \mathrm{sgn}(\mathcal{S}_0)$, with $S_0'$ and $G_0'$ defined analogously. Then*

$$\mathbb{P}\left(\mathcal{P}_{\mathcal{X}}(\tfrac{d_1d_2d_3}{n}\mathcal{P}_{\Omega_0} - I)\mathcal{P}_{\mathcal{X}} \le \sigma\right) \le \mathbb{P}\left(\mathcal{P}_{\mathcal{X}}(\tfrac{d_1d_2d_3}{n}\mathcal{P}_{\Omega_0'} - I)\mathcal{P}_{\mathcal{X}} \le \sigma\right), \tag{30}$$

$$\mathbb{P}\left(\|\mathcal{W}^{\mathcal{S}_0}\| < \frac{1}{8}\right) \le \mathbb{P}\left(\|\mathcal{W}^{\mathcal{S}_0'}\| < \frac{1}{8}\right), \tag{31}$$

*and*

$$\mathbb{P}\left(\|\mathcal{P}_{\Omega_0^\perp}\mathcal{W}^{\mathcal{S}_0}\|_{\max} < \frac{\lambda}{8}\right) \le \mathbb{P}\left(\|\mathcal{P}_{\Omega_0'^\perp}\mathcal{W}^{\mathcal{S}_0'}\|_{\max} < \frac{\lambda}{8}\right). \tag{32}$$

*Proof.* Inequality (30) is clear from Lemma 8, and this inequality implies the other two. Consider (31). The bound in (21) shows that

$$\mathbb{P}\left(\lambda\left\|\sum_{k=1}^{\infty}(\mathcal{P}_{\Omega_0}\mathcal{P}_{\mathcal{X}}\mathcal{P}_{\Omega_0})^k G_0\right\| > t\right) \le 2\exp\left(-\frac{t^2(1-\sigma^2)^2}{32\sigma^4\lambda^2} + \frac{21(d_1+d_2+d_3)}{4}\right) + \mathbb{P}(\|\mathcal{P}_{\Omega_0}\mathcal{P}_{\mathcal{X}}\| > \sigma).$$

With $\sigma$ fixed, the only dependence on $\Omega_0$ is in the second term on the right. It is clear from (30) that this probability is decreasing $|\Omega_0|$. Because $|\Omega_0'| \le |\Omega_0|$ this implies (31). The last inequality, (32), follows from the bound in (22) by the same argument.

□

Finally, we need to show that exact recovery under the random-sign model implies exact recovery under the fixed sign model.

**Lemma 16.** *Suppose $\mathcal{L}$ satisfies the conditions of Theorem 1 and that the locations of the nonzero entries of $\mathcal{S}$ follow the Bernoulli model with parameter $2\rho$. Assume further that the signs of $\mathcal{S}$ follow a Bernoulli model with parameter $\rho$. If the solution to* (6) *is exact with high probability, then it is also exact with at least the same probability for the model in which the signs are fixed and the locations are sampled from the Bernoulli model with parameter $\rho$.*

*Proof.* Lemma 16 is proved for the matrix case in [12, Thm. 2.3], and generalizing this proof to higher-orders is trivial. □

# C Proof of Lemma 6

The following lemma is implied by Lemma 1 and Theorem 1 of [43].

**Lemma 17.** *Suppose that each element of $\mathcal{X}_{i_1,i_2,i_3}$ is independent, zero-mean, and satisfies $\mathbb{E}\left[e^{t\mathcal{X}_{i_1,i_2,i_3}}\right] \le e^{k^2t^2/2}$. Then the spectral norm of $\mathcal{X}$ can be bounded as follows:*

$$\|\mathcal{X}\| \le \sqrt{8k^2\left(\sum_{i=1}^{3} d_i\right)\log(6/\log(3/2)) + \log(2/\delta)}$$

*with probability at least $1-\delta$.*

To prove Lemma 6, we just need to show that the distribution of $G$ is sub-Gaussian:

$$\mathbb{E}\left[e^{t\mathcal{X}_{i_1,i_2,i_3}}\right] = \rho_s \left( \frac{e^{t^2} + e^{-t^2}}{2} \right) \leq \cosh(t^2) \leq e^{t^2/2}.$$

# D Proof of Proposition 9

Our proof of Proposition 9 uses the following proposition:

**Proposition 18.** *Suppose $\gamma^*, \{u_r^{*(1)}\}, \cdots, \{u_r^{*(K)}\}, \mathcal{S}^*$ are optimal for the nonconvex program*

$$\begin{aligned} \min_{\gamma,\{u_r^{(1)}\},\cdots,\{u_r^{(K)}\}} \frac{1}{2} \left\| \sum_{r=1}^{R} \gamma_r (u_r^{(1)} \otimes \cdots \otimes u_r^{(K)}) + \mathcal{S} - \mathcal{Z} \right\|_F^2 + \lambda_s \|\mathcal{S}\|_{\text{sum}} \\ + \frac{\lambda_x}{K} \|\gamma\|_1 \\ \text{s.t.} \quad \|u_r^{(1)}\|, \cdots, \|u_r^{(K)}\| \leq 1, \end{aligned} \tag{33}$$

*Let $\mathcal{X}^* = \sum_{r=1}^{R} \gamma_r^* (u_r^* \otimes v_r^* \otimes w_r^*)$. Then the point $(\mathcal{X}^*, \mathcal{S}^*)$ is optimal for the problem* (24)*. Conversely, if $(\mathcal{X}^*, \mathcal{S}^*)$ is the minimizer of* (24)*, then the terms $\gamma^*, \{u_r^*\}, \{v_r^*\}, \{w_r^*\}$ from a decomposition of $\mathcal{X}^*$ are optimal for* (33).

*Proof.* Using the definition of the atomic norm, we can rewrite (6) as

$$\begin{aligned} \min_{\mathcal{X}} \quad & \frac{1}{2} \|\mathcal{X} + \mathcal{S} - \mathcal{Z}\|_F^2 + \lambda_s \|\mathcal{S}\|_{\text{sum}} + \min_{\gamma,\{u_r^{(1)}\},\cdots,\{u_r^{(K)}\}} \lambda_x \|\gamma\|_1, \\ \text{s.t.} \quad & \mathcal{X} = \sum_{r=1}^{R} \gamma_r (u_r^{(1)} \otimes \cdots \otimes u_r^{(K)}), \quad \|u_r^{(1)}\| = \cdots = \|u_r^{(K)}\| = 1. \end{aligned}$$

Due to the coerciveness of norms, replacing the norm constraints with inequalities does not change the global optima. This adjustment yields (33). □

If $R$ is chosen large enough, then the program (33) is an equivalent reformulation of (6). However, instead of replacing the atomic norm with a smooth, nonconvex regularizer as we would in the matrix case, we have introduced a non-smooth term and multiple constraints. We would like a nonconvex representation of the atomic norm that more closely generalizes (23). Proposition 9 provides this. The proof of Proposition 9 is as follows:

*Proof.* We use an argument similar to the proof in Appendix II of [9]. We can rewrite (26) as

$$\begin{aligned} \min_{\gamma,\{a_r^{(1)}\},\cdots,\{a_r^{(K)}\},\{u_r^{(1)}\},\cdots,\{u_r^{(K)}\}} \frac{1}{2} \|\mathcal{X} + \mathcal{S} - \mathcal{Z}\|_F^2 + \lambda_s \|\mathcal{S}\|_{\text{sum}} \\ + \frac{\lambda_x}{K} \sum_{r=1}^{R} \sum_{k=1}^{K} \|a_r^{(k)}\|^K \end{aligned} \tag{34}$$

$$
\begin{aligned}
\text{s.t.} \quad & \mathcal{X} = \sum_{r=1}^{R} \gamma_r (u_r^{(1)} \otimes \cdots \otimes u_r^{(K)}), \\
& \gamma_r = \|a_r^{(1)}\| \cdots \|a_r^{(K)}\|.
\end{aligned}
$$

Minimizing over $\gamma, \{a_r\}, \{b_r\}$, and $\{c_r\}$ first, we must solve

$$
\begin{aligned}
\min_{\gamma} \quad & \sum_{r=1}^{R} \sum_{k=1}^{K} \|a_r^{(k)}\|^K \\
\text{s.t.} \quad & \gamma_r = \|a_r^{(1)}\| \cdots \|a_r^{(K)}\|.
\end{aligned}
$$

The AM-GM inequality tells us

$$
(\|a_r^{(1)}\|^K \cdots \|a_r^{(K)}\|^K)^{\frac{1}{K}} \leq \tfrac{1}{K}(\|a_r^{(1)}\|^K + \cdots + \|a_r^{(K)}\|^K),
$$

with equality when $\|a_r^{(1)}\| = \cdots = \|a_r^{(K)}\| = \gamma^{\frac{1}{K}}$, so the optimal $\gamma$ satisfies

$$
\|\gamma\|_1 = \frac{1}{K} \sum_{r=1}^{R} \sum_{k=1}^{K} \|a_r^{(k)}\|^K.
$$

Using these optimal values in (34), we see that (34) is equivalent to

$$
\begin{aligned}
\min_{\gamma, \{u_r^{(1)}\}, \cdots, \{u_r^{(K)}\}} \quad & \frac{1}{2} \left\| \sum_{r=1}^{R} \gamma_r (u_r^{(1)} \otimes \cdots \otimes u_r^{(K)}) + \mathcal{S} - \mathcal{Z} \right\|_F^2 + \lambda_s \|\mathcal{S}\|_{\text{sum}} \\
& \qquad + \frac{\lambda_x}{K} \|\gamma\|_1, \\
\text{s.t.} \quad & \|u_r^{(1)}\|, \cdots, \|u_r^{(K)}\| \leq 1,
\end{aligned}
$$

which is equivalent to (6) by Proposition 18. □