
**Operations Research**

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Nonconvex Low-Rank Tensor Completion from Noisy Data

Changxiao Cai, Gen Li, H. Vincent Poor, Yuxin Chen

**To cite this article:**
Changxiao Cai, Gen Li, H. Vincent Poor, Yuxin Chen (2022) Nonconvex Low-Rank Tensor Completion from Noisy Data. Operations Research 70(2):1219-1237. https://doi.org/10.1287/opre.2021.2106

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 70, No. 2, March–April 2022, pp. 1219–1237
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

# Nonconvex Low-Rank Tensor Completion from Noisy Data

**Changxiao Cai,[a] Gen Li,[b] H. Vincent Poor,[a] Yuxin Chen[a]**

[a] Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08540; [b] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

**Contact:** ccai@princeton.edu (CC); g-li16@mails.tsinghua.edu.cn (GL); poor@princeton.edu (HVP); yuxin.chen@princeton.edu, https://orcid.org/0000-0001-9256-5815 (YC)

**Received:** November 10, 2019
**Revised:** July 26, 2020
**Accepted:** November 1, 2020
**Published Online in Articles in Advance:** June 3, 2021

**Area of Review:** Machine Learning and Data Science

**https://doi.org/10.1287/opre.2021.2106**



**Abstract:** We study a noisy tensor completion problem of broad practical interest, namely, the reconstruction of a low-rank tensor from highly incomplete and randomly corrupted observations of its entries. Whereas a variety of prior work has been dedicated to this problem, prior algorithms either are computationally too expensive for large-scale applications or come with suboptimal statistical guarantees. Focusing on "incoherent" and well-conditioned tensors of a constant canonical polyadic rank, we propose a two-stage nonconvex algorithm—(vanilla) gradient descent following a rough initialization—that achieves the best of both worlds. Specifically, the proposed nonconvex algorithm faithfully completes the tensor and retrieves all individual tensor factors within nearly linear time, while at the same time enjoying near-optimal statistical guarantees (i.e., minimal sample complexity and optimal estimation accuracy). The estimation errors are evenly spread out across all entries, thus achieving optimal $\ell_\infty$ statistical accuracy. We also discuss how to extend our approach to accommodate asymmetric tensors. The insight conveyed through our analysis of nonconvex optimization might have implications for other tensor estimation problems.



**Funding:** Y. Chen is supported in part by the Air Force Office of Scientific Research [Grant FA9550-19-1-0030], by the Office of Naval Research [Grant N00014-19-1-2120], by the Army Research Office [Grants W911NF-18-1-0303 and W911NF-20-1-0097], by the National Science Foundation (NSF) [Grants CCF-1907661, IIS-1900140, and IIS-2100158], and by a Princeton SEAS Innovation Award. H. V. Poor is supported in part by the NSF [Grant DMS-1736417]. C. Cai is supported in part by a Gordon Y. S. Wu Fellowship in Engineering. This work was done in part while Y. Chen was visiting the Kavli Institute for Theoretical Physics, supported in part by the NSF [Grant PHY-1748958].

**Supplemental Material:** The e-companion is available at https://doi.org/10.1287/opre.2021.2106.

**Keywords:** tensor completion • nonconvex optimization • gradient descent • spectral methods • entrywise statistical guarantees • minimaxity

## 1. Introduction and Motivation

### 1.1. Tensor Completion from Noisy Entries

Estimation of low-complexity models from highly incomplete observations is a fundamental task that spans a diverse array of science and engineering applications. Arguably one of the most extensively studied problems of this kind is matrix completion, where one wishes to recover a low-rank matrix given only partial entries (Davenport and Romberg 2016, Chen and Chi 2018). Moving beyond matrix-type data, a natural higher-order generalization is *low-rank tensor completion*, which aims to reconstruct a low-rank tensor when the vast majority of its entries are unseen. There is certainly no shortage of applications that motivate the investigation of tensor completion (e.g., personalized medicine (Soroushmehr and Najarian 2016, Pawlowski 2019), medical imaging (Gandy et al. 2011, Semerci et al. 2014, Cheng et al. 2017), seismic data analysis (Kreimer et al. 2013, Ely et al. 2013), and multidimensional harmonic retrieval (Chen and Chi 2014, Ying et al. 2017)). One concrete example in operations research arises when learning the preference of individual customers for a collection of products on the basis of historical transactions (Farias and Li 2019, Mišić and Perakis 2020). Given the limited availability of transaction data (e.g., each customer might only have purchased very few products before), it is crucial to exploit multiway customer-product interactions (e.g., users' browsing and searching histories) in order to better predict the likelihood of a customer purchasing a new product. Clearly, the presence of missing data and the need of exploiting a multiway structure result in the

task of tensor completion. Additionally, tensor completion finds important applications in visual data inpainting (Liu et al. 2013, Li et al. 2017), where one wishes to reconstruct video data (or a sequence of images) from incomplete measurements. The video data consist of at least two spatial variables and one temporal variable, whose intrinsic connections are often modeled via certain low-complexity tensors.

For the sake of clarity, we phrase the problem formally before we proceed, focusing on a simple model that already captures the intrinsic difficulty of tensor completion in many aspects.[1] Imagine that we are asked to estimate a symmetric order-3 tensor[2] $\boldsymbol{T}^\star \in \mathbb{R}^{d\times d\times d}$ from a small number of noisy entries

$$T_{j,k,l} = T^\star_{j,k,l} + E_{j,k,l}, \quad \forall (j,k,l) \in \Omega, \tag{1}$$

where $T_{j,k,l}$ is the observed noisy entry at location $(j,k,l)$, $E_{j,k,l}$ stands for the associated noise, and $\Omega \subseteq \{1,\cdots,d\}^3$ is a symmetric index subset to sample from. For notational simplicity, we set $\boldsymbol{T} = [T_{j,k,l}]_{1\le j,k,l\le d}$ and $\boldsymbol{E} = [E_{j,k,l}]_{1\le j,k,l\le d}$, with $T_{j,k,l} = E_{j,k,l} = 0$ for any $(j,k,l) \notin \Omega$. We adopt a *random sampling* model such that each index $(j,k,l)$ $(j \le k \le l)$ is included in $\Omega$ independently with probability $p$. In addition, we know a priori that the unknown tensor $\boldsymbol{T}^\star \in \mathbb{R}^{d\times d\times d}$ is a superposition of $r$ rank-one tensors (often termed *canonical polyadic* [CP] decomposition if $r$ is minimal):

$$\begin{aligned} \boldsymbol{T}^\star &= \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star, \quad \text{or more concisely,} \\ \boldsymbol{T}^\star &= \sum_{i=1}^{r} \boldsymbol{u}_i^{\star\otimes 3}, \end{aligned} \tag{2}$$

where each $\boldsymbol{u}_i^\star \in \mathbb{R}^d$ represents one of the $r$ low-rank tensor components/factors. Here and throughout, for any vectors $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c} \in \mathbb{R}^d$, the tensor $\boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{c}$ is a $d\times d\times d$ array whose $(j,k,l)$ th entry is given by $a_j b_k c_l$. The primary question is this: Can we hope to faithfully estimate $\boldsymbol{T}^\star$, as well as the individual tensor factors $\{\boldsymbol{u}_i^\star\}_{1\le i\le r}$, from the partially revealed entries (1), assuming that $r$ is reasonably small?

### 1.2. Computational and Statistical Challenges

Even though tensor completion conceptually resembles matrix completion in various ways, it is considerably more challenging than the matrix counterpart. This is perhaps not surprising, given that a plethora of natural tensor problems (e.g., computing the spectral norm, finding the best low-rank approximation) are all notoriously hard (Hillar and Lim 2013). As a notable example, whereas matrix completion is often efficiently solvable under nearly minimal sample complexity (Candès and Recht 2009, Gross 2011), all polynomial-time algorithms developed so far for tensor completion—even in the noise-free case—require a sample size at least exceeding the order of $rd^{3/2}$, which is substantially larger than the degrees of freedom (i.e., $rd$) underlying the model (2). In fact, it is widely conjectured that there exists a large computational barrier away from the information-theoretic sampling limits (Barak and Moitra 2016).

With this fundamental gap in mind, the current paper focuses on the regime (in terms of the sample size) that enables reliable tensor completion in polynomial time. A variety of algorithms have been proposed that enjoy some sort of theoretical guarantees in (at least part of) this regime, including, but not limited to, spectral methods (Montanari and Sun 2018, Cai et al. 2021), sum-of-squares hierarchy (Barak and Moitra 2016, Potechin and Steurer 2017), nonconvex algorithms (Jain and Oh 2014, Xia and Yuan 2017), and also convex relaxation (based on proper unfolding) (Gandy et al. 2011, Romera-Paredes and Pontil 2013, Goldfarb and Qin 2014, Huang et al. 2015). Whereas these are all polynomial-time algorithms, most of the computational complexities supported by prior theory remain prohibitively high when dealing with large-scale tensor data—a point that we shall elaborate on later. The only exception is the unfolding-based spectral method, which, however, fails to achieve exact recovery as the noise vanishes. This leads to the following critical question.

**Question 1.** *Is there any linear-time algorithm that is guaranteed to work for low-rank tensor completion?*

Going beyond such computational concerns, one might naturally wonder whether it is also possible for a fast algorithm to achieve a nearly unimprovable statistical accuracy in the presence of noise. Toward this end, intriguing stability guarantees have been established for sum-of-squares hierarchy in the noisy settings (Barak and Moitra 2016), although this paradigm is computationally expensive for large-scale data. In a recent work, Xia et al. (2017) came up with a two-stage algorithm (i.e., a spectral method followed by tensor power iterations) for noisy tensor completion. Its estimation accuracy, however, falls short of achieving exact recovery in the absence of noise. This gives rise to another question of fundamental importance.

**Question 2.** *Can we achieve near-optimal statistical accuracy without compromising computational efficiency?*

In this paper, we aim to address these two questions by developing a nonconvex algorithm that achieves optimal computational efficiency and statistical accuracy all at once.

## 2. Algorithm and Main Results

### 2.1. A Two-Stage Nonconvex Algorithm

To address the aforementioned challenges, a first impulse is to resort to the following least-squares problem:

$$\underset{\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r\in\mathbb{R}^d}{\text{minimize}} \quad \sum_{j,k,l\in\Omega} \left( \left[ \sum_{i=1}^{r} \boldsymbol{u}_i^{\otimes 3} \right]_{j,k,l} - T_{j,k,l} \right)^2, \tag{3}$$

or, more concisely (up to proper rescaling),

$$\underset{\boldsymbol{U}\in\mathbb{R}^{d\times r}}{\text{minimize}} \quad f(\boldsymbol{U}) := \frac{1}{6p}\left\|\mathcal{P}_{\Omega}\left(\sum_{i=1}^{r}\boldsymbol{u}_i^{\otimes 3}-\boldsymbol{T}\right)\right\|_{\mathrm{F}}^2, \tag{4}$$

if we take $\boldsymbol{U} := [\boldsymbol{u}_1,\ldots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}$. Here, we denote by $\mathcal{P}_{\Omega}(\boldsymbol{T})$ the orthogonal projection of any tensor $\boldsymbol{T}$ onto the subspace of tensors that vanish outside of the index set $\Omega$. This optimization problem, however, is highly nonconvex (which involves minimizing a degree-6 polynomial), thus resulting in computational intractability in general.

Fortunately, not all nonconvex problems are as daunting to solve as they may seem. For example, recent years have seen a flurry of activity in low-rank matrix factorization via nonconvex optimization, which provably achieves optimal statistical accuracy and computational efficiency at once (see Chi et al. 2019 for an overview of recent advances). Motivated by this strand of work, we propose to solve (4) via a two-stage nonconvex paradigm, which we present in reverse order. The whole procedure is summarized in Algorithms 1–3.

**2.1.1. Gradient Descent (GD).** Arguably one of the simplest optimization algorithms is gradient descent, which adopts a gradient update rule

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^{t} - \eta_t\nabla f(\boldsymbol{U}^{t}), \quad t = 0, 1, \cdots, \tag{5}$$

where $\eta_t$ is the learning rate or the step size and $\boldsymbol{U}^t\in\mathbb{R}^{d\times r}$ is the estimate in the $t$ th iteration. The main computational burden in each iteration lies in gradient evaluation, which, in this case, can be performed in time proportional to that taken to read the data.

Despite the simplicity of this algorithm, two critical issues stand out and might significantly affect its efficiency, which we shall bear in mind throughout the algorithmic and theoretical development.

i. *Local stationary points and initialization.* As is well known, GD is guaranteed to find an approximate local stationary point, provided that the learning rates do not exceed the inverse Lipschitz constant of the gradient (Bubeck 2015). There exist, however, local stationary points (e.g., saddle points or spurious local minima) that might fall short of the desired statistical properties. This requires us to properly avoid such undesired points, while retaining computational efficiency. To address this issue, one strategy is to first identify a rough initial guess within a local region surrounding the global solution (which often helps rule out bad local minima), in order to guarantee proper convergence of subsequent optimization procedures (Jain and Oh 2014, Li and Tang 2017). As a side remark, although careful initialization might not be crucial for several matrix recovery cases (Gilboa et al. 2018, Chen et al. 2019c, Tan and Vershynin 2019), it does seem to be critical in various tensor problems (Richard and Montanari 2014). We shall elucidate this point in Section EC.1 in the e-companion.

ii. *Learning rates and regularization.* Learning rates play a pivotal role in determining the convergence properties of GD. The challenge, however, is that the loss function (4) is overall not sufficiently smooth (i.e., its gradient often has an exceedingly large Lipschitz constant), and hence generic optimization theory recommends a pessimistically slow update rule (i.e., an extremely small learning rate) so as to guard against overshooting. This, however, slows down the algorithm significantly, thus destroying the main computational advantage of GD (i.e., low per-iteration cost). With this issue in mind, prior literature suggests carefully designed regularization steps (e.g., proper projection, regularized loss functions) in order to improve the geometry of the optimization landscape (Xia and Yuan 2017). In contrast, we argue that one is allowed to take a constant learning rate—which is as aggressive as it can possibly be—even without enforcing any regularization procedures.

**Algorithm 1** (Gradient Descent for Nonconvex Tensor Completion)

1: Generate an initial estimate $\boldsymbol{U}^0\in\mathbb{R}^{d\times r}$ via Algorithm 2.

2: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

3: $\boldsymbol{U}^{t+1} = \boldsymbol{U}^{t} - \eta_t\nabla f(\boldsymbol{U}^{t}) = \boldsymbol{U}^{t} - \frac{\eta_t}{p}\mathcal{P}_{\Omega}(\sum_{i=1}^{r}(\boldsymbol{u}_i^t)^{\otimes 3} - \boldsymbol{T})\times_1^{\mathsf{seq}}\boldsymbol{U}^t\times_2^{\mathsf{seq}}\boldsymbol{U}^t$, where $\times_1^{\mathsf{seq}}$ and $\times_2^{\mathsf{seq}}$ are defined in Section 2.4.

**2.1.2. Initialization.** Motivated by the aforementioned issue (i), we develop a procedure that guarantees a reasonable initial estimate. In a nutshell, the proposed procedure consists of two steps:

a. estimate the subspace spanned by the $r$ low-rank tensor factors $\{\boldsymbol{u}_i^\star\}_{1\le i\le r}$ via a spectral method;

b. disentangle individual low-rank tensor factors from this subspace estimate.

As we shall see momentarily, the total computational complexity of the proposed initialization is $O(pd^3)$ when $r = O(1)$, $\kappa = O(1)$, and $p\ge 1/d^2$ (where $\kappa$ is a sort of "condition number" defined later), which is a linear-time algorithm. Note, however, that these two steps in the initialization procedure are relatively more complicated to describe. To improve the flow of the current paper, we postpone the details to Section 3. The readers can catch a glimpse of these procedures in Algorithms 2–3.

**Algorithm 2** (Spectral Initialization for Nonconvex Tensor Completion)

1: Let $\boldsymbol{U}\Lambda\boldsymbol{U}^\top$ be the rank-$r$ eigen-decomposition of

$$\boldsymbol{B} := \mathcal{P}_{\mathsf{off-diag}}(\boldsymbol{A}\boldsymbol{A}^\top), \tag{6}$$

where $\boldsymbol{A} = \mathsf{unfold}(p^{-1}\boldsymbol{T})$ is the mode-1 matricization of $p^{-1}\boldsymbol{T}$ and $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{Z})$ extracts out the off-diagonal entries of $\boldsymbol{Z}$.

2: **Output:** an initial estimate $\boldsymbol{U}^0\in\mathbb{R}^{d\times r}$ on the basis of $\boldsymbol{U}\in\mathbb{R}^{d\times r}$ using Algorithm 3.

**Algorithm 3** (Retrieval of Low-Rank Tensor Factors from a Given Subspace Estimate)

1: **Input:** number of restarts $L$, pruning threshold $\epsilon_{\text{th}}$, subspace estimate $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ given by Algorithm 2.

2: **for** $\tau = 1, \ldots, L$ **do**

3: Generate an independent Gaussian vector $\boldsymbol{g}^\tau \sim \mathcal{N}(0, \boldsymbol{I}_d)$.

4: $(\boldsymbol{\nu}^\tau, \lambda_\tau, \text{spec-gap}_\tau) \leftarrow$ RETRIEVE-ONE-TENSOR-FACTOR $(\boldsymbol{T}, p, \boldsymbol{U}, \boldsymbol{g}^\tau)$.

5: Generate tensor factor estimates $\{(\boldsymbol{w}^1, \lambda_1), \ldots, (\boldsymbol{w}^r, \lambda_r)\} \leftarrow$ PRUNE$(\{(\boldsymbol{\nu}^\tau, \lambda_\tau, \text{spec-gap}_\tau)\}_{\tau=1}^L, \epsilon_{\text{th}})$.

6: **Output:** initial estimate $\boldsymbol{U}^0 = [\lambda_1^{1/3}\boldsymbol{w}^1, \ldots, \lambda_r^{1/3}\boldsymbol{w}^r]$.

1: **function** RETRIEVE-ONE-TENSOR-FACTOR$(\boldsymbol{T}, p, \boldsymbol{U}, \boldsymbol{g})$

2: Compute

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{g} =: \mathcal{P}_{\boldsymbol{U}}(\boldsymbol{g}), \tag{7a}$$

$$\boldsymbol{M} = p^{-1}\boldsymbol{T}\times_3 \boldsymbol{\theta}, \tag{7b}$$

where $\times_3$ is defined in Section 2.4.

3: Let $\boldsymbol{\nu}$ be the leading singular vector of $\boldsymbol{M}$ obeying $\langle \boldsymbol{T}, \boldsymbol{\nu}^{\otimes 3}\rangle \geq 0$, and set $\lambda = \langle p^{-1}\boldsymbol{T}, \boldsymbol{\nu}^{\otimes 3}\rangle$.

4: **return** $(\boldsymbol{\nu}, \lambda, \sigma_1(\boldsymbol{M}) - \sigma_2(\boldsymbol{M}))$.

1: **function** PRUNE$(\{(\boldsymbol{\nu}^\tau, \lambda_\tau, \text{spec}-\text{gap}_\tau)\}_{\tau=1}^L, \epsilon_{\text{th}})$

2: Set $\Theta = \{(\boldsymbol{\nu}^\tau, \lambda_\tau, \text{spec}-\text{gap}_\tau)\}_{\tau=1}^L$.

3: **for** $i = 1, \ldots, r$ **do**

4: Choose $(\boldsymbol{\nu}^\tau, \lambda_\tau, \text{spec}-\text{gap}_\tau)$ from $\Theta$ with the largest $\text{spec}-\text{gap}_\tau$; set $\boldsymbol{w}^i = \boldsymbol{\nu}^\tau$ and $\lambda_i = \lambda_\tau$.

5: Update $\Theta \leftarrow \Theta \setminus \{(\boldsymbol{\nu}^\tau, \lambda_\tau, \text{spec-gap}_\tau) \in \Theta : |\langle \boldsymbol{\nu}^\tau, \boldsymbol{w}^i\rangle| > 1 - \epsilon_{\text{th}}\}$.

6: **return** $\{(\boldsymbol{w}^1, \lambda_1), \ldots, (\boldsymbol{w}^r, \lambda_r)\}$.

## 2.2. Main Results

Encouragingly, the proposed nonconvex algorithm provably achieves the best of both worlds—in terms of statistical accuracy and computational efficiency—for a class of low-rank, well-conditioned, and "incoherent" problem instances. This subsection summarizes our main findings.

Before continuing, we note that one cannot hope to recover an arbitrary tensor from highly subsampled and arbitrarily corrupted entries. In order to enable provably valid recovery, the present paper focuses on a tractable model by imposing the following assumptions.

**Definition 1** (Incoherence and Well-Conditionedness). Define the incoherence parameters and the condition number of $\boldsymbol{T}^\star$ as follows:

$$\mu_0 := \frac{d^3\|\boldsymbol{T}^\star\|_\infty^2}{\|\boldsymbol{T}^\star\|_{\text{F}}^2}, \tag{8a}$$

$$\mu_1 := \frac{d\|\boldsymbol{u}_i^\star\|_\infty^2}{\|\boldsymbol{u}_i^\star\|_2^2}, \tag{8b}$$

$$\mu_2 := \frac{d\langle \boldsymbol{u}_i^\star, \boldsymbol{u}_j^\star\rangle^2}{\|\boldsymbol{u}_i^\star\|_2^2\|\boldsymbol{u}_j^\star\|_2^2}, \tag{8c}$$

$$\kappa := \frac{\max_i \|\boldsymbol{u}_i^\star\|_2}{\min_i \|\boldsymbol{u}_i^\star\|_2}. \tag{8d}$$

**Remark 1.** Here, $\mu_0$, $\mu_1$, and $\mu_2$ are termed the *incoherence parameters*. Definitions (8a)–(8c) can be viewed as some sort of incoherence conditions for the tensor. For instance, when $\mu_0, \mu_1$, and $\mu_2$ are small, these conditions say that (1) the energy of tensor $\boldsymbol{T}^\star$ is (nearly) evenly spread across all entries; (2) each factor $\boldsymbol{u}_i^\star$ is delocalized; (3) the factors $\{\boldsymbol{u}_i^\star\}$ are nearly orthogonal to each other. Definition (8d) is concerned with the "well-conditionedness" of the tensor, meaning that each rank-1 component is of roughly the same size. In particular, we note that an assumption on pairwise correlation (i.e., a constraint on $\mu_2$) is often assumed in the literature of tensor decomposition/factorization (e.g., Anandkumar et al. 2014b, Sun et al. 2017, Hao et al. 2020).

For notational simplicity, we shall set

$$\mu := \max\{\mu_0, \mu_1, \mu_2\}. \tag{9}$$

Note that our theory allows $\mu$ to grow with the problem dimension $d$ (in fact, $\mu$ can be as large as $d/\text{polylog}(d)$).

**Assumption 1** (Random Noise). *Suppose that $\boldsymbol{E}$ is a symmetric random tensor, where $\{E_{j,k,l}\}_{1\leq j\leq k\leq l\leq d}$ (see (1)) are independently generated sub-Gaussian random variables with mean zero and variance* $\text{Var}(E_{j,k,l}) \leq \sigma^2$.

In addition, recognizing that there is a global permutational ambiguity issue (i.e., one cannot distinguish $\boldsymbol{u}_1^\star, \cdots, \boldsymbol{u}_r^\star$ from an arbitrary permutation of them), we introduce the following loss metrics to account for this ambiguity:

$$\text{dist}_{\text{F}}(\boldsymbol{U}, \boldsymbol{U}^\star) := \min_{\Pi\in\text{perm}_r} \|\boldsymbol{U}\Pi - \boldsymbol{U}^\star\|_{\text{F}}, \tag{10a}$$

$$\text{dist}_{\infty}(\boldsymbol{U}, \boldsymbol{U}^\star) := \min_{\Pi\in\text{perm}_r} \|\boldsymbol{U}\Pi - \boldsymbol{U}^\star\|_{\infty}, \tag{10b}$$

$$\text{dist}_{2,\infty}(\boldsymbol{U}, \boldsymbol{U}^\star) := \min_{\Pi\in\text{perm}_r} \|\boldsymbol{U}\Pi - \boldsymbol{U}^\star\|_{2,\infty}, \tag{10c}$$

where $\text{perm}_r$ stands for the set of $r\times r$ permutation matrices. For notational simplicity, we also take

$$\lambda_{\min}^\star := \min_{1\leq i\leq r} \|\boldsymbol{u}_i^\star\|_2^3 \quad \text{and} \quad \lambda_{\max}^\star := \max_{1\leq i\leq r} \|\boldsymbol{u}_i^\star\|_2^3. \tag{11}$$

With these notations in place, we are ready to present our main results. For simplicity of presentation, we shall start with the setting where $r, \mu, \kappa \asymp 1$.

**Theorem 1.** *Fix an arbitrary small constant $\delta > 0$. Suppose that $r, \kappa, \mu = O(1)$,*

$$p \geq c_0 \frac{\log^4 d}{d^{3/2}}, \qquad \frac{\sigma}{\lambda_{\min}^\star} \leq c_1 \frac{\sqrt{p}}{d^{3/4}\log^2 d},$$

$$L = c_2 \quad \text{and} \quad \epsilon_{\text{th}} = c_3\left(\frac{\log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log^2 d}{p}} + \sqrt{\frac{\log d}{d}}\right)$$

*for some sufficiently large constants $c_0, c_2 > 0$ and some sufficiently small constants $c_1, c_3 > 0$. The learning rate $\eta_t \equiv \eta$ is*

*taken to be a constant obeying* $0 < \eta \le \lambda_{\min}^{\star 4/3}/(32\lambda_{\max}^{\star 8/3})$. *Then with probability at least* $1-\delta$,

$$\mathrm{dist}_{\mathrm{F}}(\boldsymbol{U}^t, \boldsymbol{U}^\star) \le \left( C_1\rho^t + C_2 \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \right) \| \boldsymbol{U}^\star \|_{\mathrm{F}}, \quad (12a)$$

$$\mathrm{dist}_{\infty}(\boldsymbol{U}^t, \boldsymbol{U}^\star) \le \mathrm{dist}_{2,\infty}(\boldsymbol{U}^t, \boldsymbol{U}^\star) \le \left( C_3\rho^t + C_4 \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \right) \| \boldsymbol{U}^\star \|_{2,\infty} \quad (12b)$$

*hold simultaneously for all* $0 \le t \le t_0 = d^5$. *Here,* $0 < C_1, C_3, \rho < 1$ *and* $C_2, C_4 > 0$ *are some absolute constants.*

**Remark 2.** The theorem holds unchanged if $d^5$ is replaced by $d^c$ for an arbitrarily large constant $c > 0$.

**Remark 3.** The upper bound $t_0$ on the iteration count arises from the leave-one-out analysis when handling noisy observations. In short, the leave-one-out argument can only provide high-probability bounds for each iteration, thus requiring an upper bound on the iteration count if we desire a uniform bound across iterations. Note that in the noiseless case, our results and analysis hold for an arbitrarily large number of iterations.

As an immediate consequence of Theorem 1, we obtain appealing $\ell_\infty$ statistical guarantees for estimating tensor entries, which are previously rarely available (see Table 1). Specifically, let our tensor estimate in the $t$ th iteration be

$$\boldsymbol{T}^t := \sum_{i=1}^{r} \boldsymbol{u}_i^t \otimes \boldsymbol{u}_i^t \otimes \boldsymbol{u}_i^t, \quad \text{where } \boldsymbol{U}^t = [\boldsymbol{u}_1^t, \cdots, \boldsymbol{u}_r^t] \in \mathbb{R}^{d\times r}. \quad (13)$$

Then our result is the following.

**Corollary 1.** *Fix an arbitrarily small constant* $\delta > 0$. *Instate the assumptions of* Theorem 1. *Then with probability at least* $1-\delta$,

$$\| \boldsymbol{T}^t - \boldsymbol{T}^\star \|_{\mathrm{F}} \lesssim \left( C_1\rho^t + C_2 \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \right) \| \boldsymbol{T}^\star \|_{\mathrm{F}}, \quad (14a)$$

$$\| \boldsymbol{T}^t - \boldsymbol{T}^\star \|_{\infty} \lesssim \left( C_3\rho^t + C_4 \frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}} \right) \| \boldsymbol{T}^\star \|_{\infty} \quad (14b)$$

*hold simultaneously for all* $0 \le t \le t_0 = d^5$. *Here,* $0 < C_1, C_3, \rho < 1$ *and* $C_2, C_4 > 0$ *are some absolute constants.*

Several important implications are provided as follows. The following discussion assumes that $\lambda_{\max}^\star \asymp \lambda_{\min}^\star \asymp 1$ for notational simplicity.

1. *Linear convergence.* In the absence of noise, the proposed algorithm converges linearly; namely, it provably attains $\varepsilon$ accuracy within $O(\log(1/\varepsilon))$ iterations. Given the inexpensiveness of each gradient iteration, this algorithm can be viewed as a linear-time algorithm, which can almost be implemented as long as we can read the data. In the noisy setting, the algorithm reaches an appealing statistical accuracy within a logarithmic number of iterations.

2. *Near-optimal sample complexity.* The fast convergence is guaranteed as soon as the sample size exceeds the order of $d^{3/2}\mathrm{poly}\log d$. This matches the minimal sample complexity—modulo some logarithmic factor—known so far for any polynomial-time algorithm.

3. *Near-optimal statistical accuracy.* The proposed algorithm converges geometrically fast to a point with Euclidean error $O(\sigma\sqrt{(d\log d)/p})$. This matches the lower bound established in (Xia et al. 2017, theorem 5) up to some logarithmic factor, thus justifying the statistical optimality of the proposed nonconvex algorithm.

4. *Entrywise estimation accuracy.* In addition to the Euclidean statistical guarantees, we have also established an entrywise error bound, which, to the best of our knowledge, has not been established in any of the prior work. When $t$ is sufficiently large, the iterates reach an entrywise error bound $O(\sigma\sqrt{(\log d)/p})$. This entrywise error bound is about an order of $\sqrt{d}$ times smaller than the above $\ell_2$ error bound, thereby implying that the estimation errors are evenly spread out across all entries.

5. *Noise size.* The aforementioned theory operates in the regime where $\sigma \lesssim \sqrt{\frac{p}{d^{3/2}}}$ (modulo some log factor). Given that we have $\|\boldsymbol{T}^\star\|_\infty \asymp d^{-3/2}$ in this case, our noise size constraint can be equivalently written as (up to some log factor)

$$\frac{\sigma}{\|\boldsymbol{T}^\star\|_\infty} \lesssim \sqrt{pd^{3/2}}. \quad (15)$$

Since the sampling rate needs to satisfy $p \gg d^{-3/2}$, this condition essentially allows the typical size of each noise component to be considerably larger than the size of the corresponding entry of the truth, which covers a broad range of practical scenarios.

6. *Implicit regularization.* One appealing feature of our finding is the simplicity of the iterative refinement stage of the algorithm. All of the aforementioned statistical and computational benefits hold for vanilla gradient descent (when properly initialized). This should be contrasted with prior work (e.g., Xia and Yuan 2017) that relies on extra regularization terms to stabilize the optimization landscape. In principle, vanilla gradient descent implicitly constrains itself within a region of well-conditioned landscape, thus enabling fast convergence without explicit regularization.

7. *No need of sample splitting.* The theory developed herein does not require fresh samples in each iteration. We note that sample splitting has been frequently adopted in other context primarily to simplify mathematical analysis. Nevertheless, it typically does not exploit the data in an efficient manner (i.e., each data sample is used only once), thus resulting in the need of a much larger sample size in practice.

**Table 1.** Comparison with Prior Theory for Existing Methods When $r, \mu, \kappa \asymp 1$ (Neglecting Logarithmic Factors)

| | Algorithm | Sample complexity | Computational complexity | $l_2$ error (noisy) | $l_\infty$ error (noisy) | Recovery type (noiseless) |
|---|---|---|---|---|---|---|
| Our theory | Spectral method + (vanilla) GD | $d^{1.5}$ | $pd^3$ | $\sigma\sqrt{\frac{d}{p}}$ | $\sigma\sqrt{\frac{1}{p}}$ | Exact |
| Xia et al. (2017) | Spectral initialization + tensor power method | $d^{1.5}$ | $pd^3$ | $(\|\|\boldsymbol{T}^\star\|\|_\infty + \sigma)\sqrt{\frac{d}{p}}$ | N/A | Approximate |
| Xia and Yuan (2017) | Spectral method + GD on manifold | $d^{1.5}$ | poly($d$) | N/A | N/A | Exact |
| Montanari and Sun (2018) | Spectral method | $d^{1.5}$ | $d^3$ | N/A | N/A | Approximate |
| Barak and Moitra (2016) | Sum-of-squares | $d^{1.5}$ | $d^{15}$ | $\frac{\|\|\boldsymbol{T}^\star\|\|_F}{\sqrt{pd^{1.5}}} + \sigma d^{1.5}$ | N/A | Approximate |
| Potechin and Steurer (2017) | Sum-of-squares | $d^{1.5}$ | $d^{10}$ | N/A | N/A | Exact |
| Yuan and Zhang (2016) | Tensor nuclear norm | $d$ | NP-hard | N/A | N/A | Exact |
| Yuan and Zhang (2017) | Minimization | | | | | |

*Note.* N/A, not applicable.

We shall take a moment to discuss the merits of our approach in comparison with prior work. One of the best-known polynomial-time algorithms is the degree-6 level of the sum-of-squares (SoS) hierarchy, which seems to match the computationally feasible limit in terms of the sample complexity (Barak and Moitra 2016). However, this approach has a well-documented limitation in that it involves solving a semidefinite program of dimensions $d^3 \times d^3$, which requires enormous storage and computation power. The work of Montanari and Sun (2018) alleviates this computational burden by resorting to a clever unfolding-based spectral algorithm; it is a nearly linear-time procedure that enables near-minimal sample complexity (among polynomial-time algorithms), although it does not achieve exact recovery even in the absence of noise. The two-stage algorithm developed by Xia et al. (2017)—which is based on spectral initialization followed by tensor power methods—shares similar advantages and drawbacks as the method developed by Montanari and Sun (2018). Further, the recent work of Xia and Yuan (2017) proposes a polynomial-time nonconvex algorithm based on gradient descent over a Grassmann manifold (with a properly regularized objective function), which is an extension of the nonconvex matrix completion algorithm proposed by (Keshavan et al. 2010a,b) to tensor data. The theory provided by Xia and Yuan (2017), however, does not provide explicit computational complexities. The recent work of Shah and Yu (2019) attempts tensor estimation via an interesting algorithm adapted from collaborative filtering and investigates both $\ell_2$ and $\ell_\infty$ estimation accuracy. This approach, however, does not guarantee exact recovery in the absence of noise. We summarize and compare several prior results in Table 1 (omitting logarithmic factors).

Thus far, we have concentrated on the low-rank, well-conditioned, and incoherent case. Our main theory can be extended to cover a broader class of scenarios.

**Theorem 2.** *Fix an arbitrary small constant $\delta > 0$. Suppose that $\kappa \asymp 1$,*

$$p \geq c_0 \frac{\mu^4 r^4 \log^4 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda^\star_{\min}} \leq c_1 \frac{\sqrt{p}}{\mu r^{3/2} d^{3/4} \log^2 d},$$
$$r \leq c_2 \left(\frac{d}{\mu^6 \log^6 d}\right)^{1/6},$$
$$L = c_3 r^{2\kappa^2} \log^{3/2} r \quad \text{and}$$
$$\epsilon_{\text{th}} = c_4 \left( \frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{rd \log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}} \right)$$

*for some sufficiently large constants $c_0, c_3 > 0$ and some sufficiently small constants $c_1, c_2, c_4 > 0$. The learning rate $\eta_t \equiv \eta$ is taken to be a constant obeying $0 < \eta \leq \lambda^{\star 4/3}_{\min} / (32 \lambda^{\star 8/3}_{\max})$. Then with probability at least $1 - \delta$,*

$$\text{dist}_F(\boldsymbol{U}^t, \boldsymbol{U}^\star) \leq \left( C_1 \rho^t + C_2 \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{d \log d}{p}} \right) \| \boldsymbol{U}^\star \|_F, \quad (16a)$$

$$\text{dist}_\infty(\boldsymbol{U}^t, \boldsymbol{U}^\star) \leq \text{dist}_{2,\infty}(\boldsymbol{U}^t, \boldsymbol{U}^\star) \leq \left( C_3 \rho^t + C_4 \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{d \log d}{p}} \right) \| \boldsymbol{U}^\star \|_{2,\infty} \quad (16b)$$

*hold simultaneously for all $0 \leq t \leq t_0 = d^5$. Here, $0 < C_1, C_3, \rho < 1$ and $C_2, C_4 > 0$ are some absolute constants.*

**Corollary 2.** *Fix an arbitrarily small constant $\delta > 0$. Instate the assumptions of* Theorem 2. *Then with probability at least $1-\delta$,*

$$\|\boldsymbol{T}^t - \boldsymbol{T}^\star\|_{\mathrm{F}} \lesssim \left(C_1\rho^t + C_2\frac{\sigma}{\lambda^\star_{\min}}\sqrt{\frac{d\log d}{p}}\right)\|\boldsymbol{T}^\star\|_{\mathrm{F}}, \quad (17a)$$

$$\|\boldsymbol{T}^t - \boldsymbol{T}^\star\|_{\infty} \lesssim \left(C_3\rho^t + C_4\frac{\sigma}{\lambda^\star_{\min}}\sqrt{\frac{\mu^3 r d\log d}{p}}\right)\|\boldsymbol{T}^\star\|_{\infty} \quad (17b)$$

*hold simultaneously for all $0 \le t \le t_0 = d^5$. Here, $0 < C_1, C_3, \rho < 1$ and $C_2, C_4 > 0$ are some absolute constants.*

**Remark 4.** Clearly, Theorem 2 and Corollary 2 subsume Theorem 1 and Corollary 1, respectively, as special cases.

**Remark 5.** Our theorems require the rank $r$ to not exceed $o(d^{1/6})$, which, we believe, is an artifact of the current nonconvex analysis (particularly for the initialization stage). For instance, our local convergence analysis is built upon strong convexity and smoothness, which holds only within a sufficiently small neighborhood surrounding the truth; given that the diameter of this neighborhood is no more than $o(1/r)$, our analysis requires an initial guess with higher accuracy than expected, thus leading to our rank constraint. It might be possible to improve the rank dependency via more refined analysis, and we leave it to future investigation.

In a nutshell, this theorem reveals intriguing theoretical support (including both $\ell_{\mathrm{F}}$ and $\ell_{2,\infty}$ bounds) for more general settings. Assuming that the condition number $\kappa \asymp 1$, the nonconvex algorithm that we propose is guaranteed to succeed in polynomial time. Note, however, that our theoretical dependency (including both sample and computational complexities) on the rank $r$ and the incoherence parameter $\mu$ are likely loose and suboptimal. In addition, if $\kappa$ is allowed to grow with $d$, then the current theory requires a large number of restart attempts during the initialization stage, resulting in a very high computational burden. Improving these aspects, however, calls for a much more refined analysis framework, which we leave for future investigation.

### 2.3. Numerical Experiments

We carry out a series of numerical experiments to corroborate our theoretical findings. Before proceeding, recall that Theorem 2 only guarantees successful recovery with probability $1-\delta$ for some small constant $\delta$; this means that we shall not anticipate a very high success rate (e.g., $1 - O(d^{-5})$), as in the matrix recovery case. As we shall make clear shortly, this happens mainly because the initialization stage works only with probability $1-\delta$, where the uncertainty largely depends on the random vectors $\{\boldsymbol{g}^\tau\}_{1\le\tau\le L}$. With this observation in mind, we recommend the following modification to improve the empirical success rate:

- Run Algorithm 2 independently for $t_{\mathrm{init}} = 5$ times to obtain multiple initial estimates (denoted by $\boldsymbol{U}^0_{[1]}, \cdots, \boldsymbol{U}^0_{[t_{\mathrm{init}}]}$); select the one achieving the smallest empirical loss, namely,

$$\boldsymbol{U}^0_{\mathrm{best}} = \underset{\boldsymbol{U}\in\{\boldsymbol{U}^0_{[i]}\}_{1\le i\le t_{\mathrm{init}}}}{\arg\min} f(\boldsymbol{U}). \quad (18)$$

- Run Algorithm 1 with the initial point $\boldsymbol{U}^0$ set to be $\boldsymbol{U}^0_{\mathrm{best}}$.

The final estimates for the low-rank factor and the whole tensor are denoted, respectively, by

$$\hat{\boldsymbol{U}} = \boldsymbol{U}^{t_0} \quad \text{and} \quad \hat{\boldsymbol{T}} = \sum_{i=1}^{r} \boldsymbol{u}_i^{t_0}\otimes\boldsymbol{u}_i^{t_0}\otimes\boldsymbol{u}_i^{t_0}, \quad (19)$$

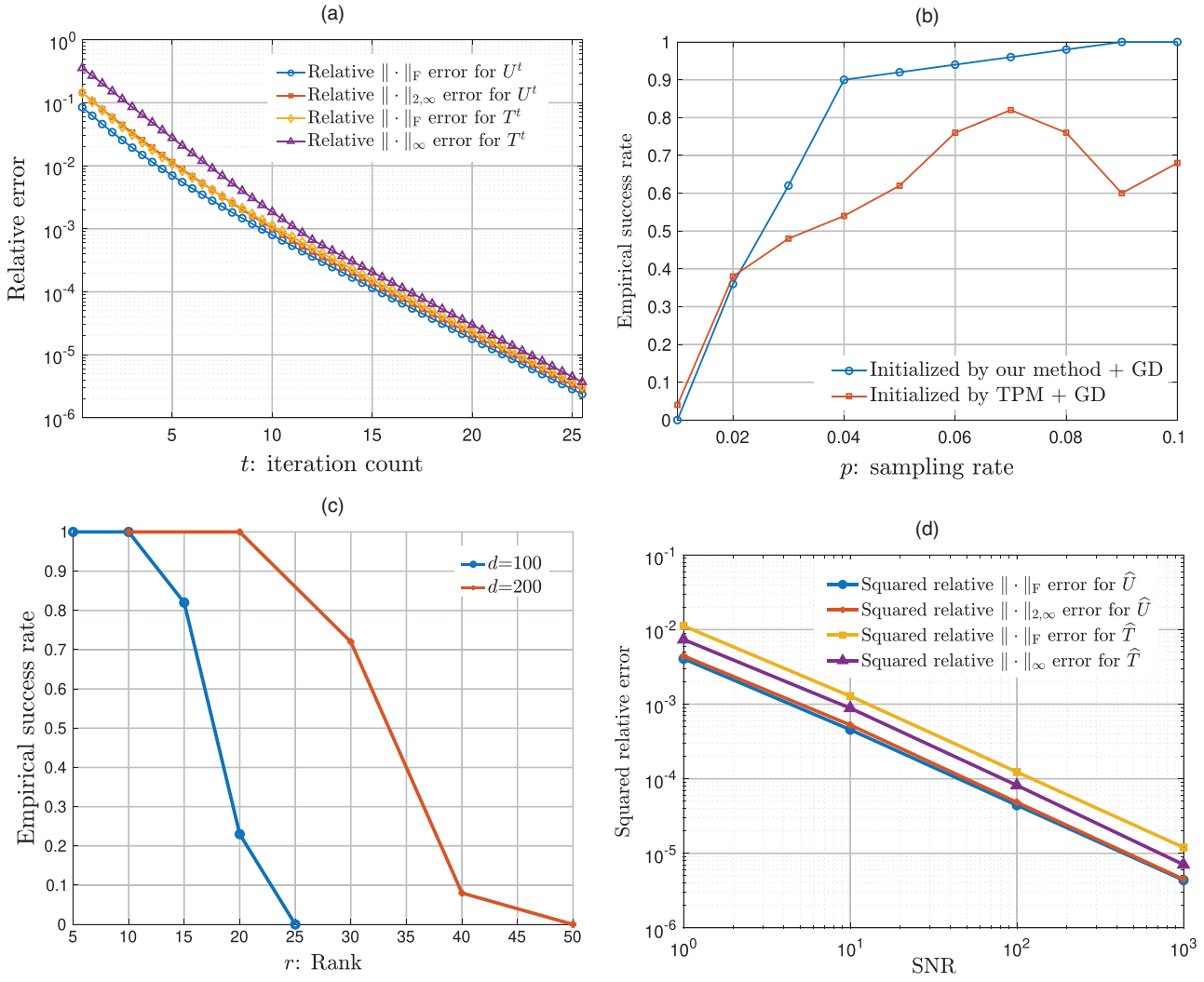
where $\boldsymbol{U}^{t_0} = [\boldsymbol{u}_1^{t_0}, \cdots, \boldsymbol{u}_r^{t_0}] \in \mathbb{R}^{d\times r}$ is the iterate returned by Algorithm 1, with $t_0$ the total number of gradient iterations. In the sequel, we generate the true tensor $\boldsymbol{T}^\star = \sum_{1\le i\le r}\boldsymbol{u}_i^{\star\otimes 3}$ randomly in such a way that $\boldsymbol{u}_i^\star \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_d)$. The learning rates are taken to be $\eta_t \equiv 0.2$ unless otherwise noted.

We start with numerical convergence rates of our algorithm in the absence of noise. Set $d = 100$, $r = 4$, $p = 0.1$, $L = 16$, and $\epsilon_{\mathrm{th}} = 0.4$. Figure 1(a) shows the numerical estimation errors versus iteration count $t$ in a typical Monte Carlo trial. Here, four kinds of estimation errors are reported: (1) the relative Frobenius norm error $\frac{\mathrm{dist}_{\mathrm{F}}(\boldsymbol{U}^t,\boldsymbol{U}^\star)}{\|\boldsymbol{U}^\star\|_{\mathrm{F}}}$; (2) the relative $\|\cdot\|_{2,\infty}$ error $\frac{\mathrm{dist}_{2,\infty}(\boldsymbol{U}^t,\boldsymbol{U}^\star)}{\|\boldsymbol{U}^\star\|_{2,\infty}}$; (3) the relative Frobenius norm error $\frac{\|\boldsymbol{T}^t-\boldsymbol{T}^\star\|_{\mathrm{F}}}{\|\boldsymbol{T}^\star\|_{\mathrm{F}}}$; and (4) the relative $\ell_\infty$ error $\frac{\|\boldsymbol{T}^t-\boldsymbol{T}^\star\|_\infty}{\|\boldsymbol{T}^\star\|_\infty}$. Here, $\boldsymbol{T}^t = \sum_{i=1}^r \boldsymbol{u}_i^t\otimes\boldsymbol{u}_i^t\otimes\boldsymbol{u}_i^t$ with $\boldsymbol{U}^t = [\boldsymbol{u}_1^t, \cdots, \boldsymbol{u}_r^t]$. For all these metrics, the numerical estimation errors decay geometrically fast.

Next, we study the phase transition (in terms of the success rates for exact recovery) in the noise-free settings. Set $d = 100$, $r = 4$, $L = 16$, and $\epsilon_{\mathrm{th}} = 0.4$. For the sake of comparisons, we also report the numerical performance of the tensor power method (TPM) followed by gradient descent. When running the tensor power method, we set both the number of iterations and the restart number to be 16. Each trial is claimed to succeed if the relative $\ell_2$ error obeys $\frac{\mathrm{dist}_{\mathrm{F}}(\hat{\boldsymbol{U}},\boldsymbol{U}^\star)}{\|\boldsymbol{U}^\star\|_{\mathrm{F}}} \le 0.01$. Figure 1(b) plots the empirical success rates over 100 independent Monte Carlo trials. As can be seen, our initialization algorithm outperforms the tensor power method.

The third series of experiments is concerned with the dependence of the success rate on the rank $r$. Let us set $p = rd^{-3/2}\log^2 d$, $L = r^2$, and $\epsilon_{\mathrm{th}} = 0.4$, and the success recovery criterion is the same as mentioned earlier. Figure 1(c) depicts the empirical success rates (over 100 independent Monte Carlo trials) as the rank $r$ varies. As can be seen from the plots, the proposed

**Figure 1.** (Color online) (a) Relative Errors of the Estimates $\boldsymbol{U}^t$ and $\boldsymbol{T}^t$ vs. Iteration Count $t$ for Noiseless Tensor Completion, Where $d = 100$, $r = 4$, and $p = 0.1$; (b) Empirical Success Rate vs. Sampling Rate, Where $d = 100$ and $r = 4$; (c) Empirical Success Rate vs. Rank, Where $p = rd^{-3/2}\log^2 d$; (d) Squared Relative Errors vs. SNR for Noisy Settings, Where $d = 100$, $r = 4$, and $p = 0.1$




*Note.* Each point in (b), (c) and (d) is averaged over 100 independent Monte Carlo trials.

algorithm is able to achieve exact reconstruction as long as the rank $r$ is sufficiently small compared with $d$. The plausible range of $r$, however, seems to be larger than our theoretic requirement $r = o(d^{1/6})$. This, once again, suggests the need of future investigation to pin down the best possible dependency on $r$.

Finally, we consider the numerical estimation accuracy of our algorithm. Take $t_0 = 100$, $d = 100$, $r = 4$, $p = 0.1$, $L = 16$, and $\epsilon_{\text{th}} = 0.4$. Define the signal-to-noise ratio (SNR) to be $\text{SNR} = \frac{\|\boldsymbol{T}^\star\|_{\text{F}}^2/d^3}{\sigma^2}$. We report in Figure 1(d) three types of squared relative errors (i.e., $\frac{\text{dist}_{\text{F}}^2(\hat{\boldsymbol{U}},\boldsymbol{U}^\star)}{\|\boldsymbol{U}^\star\|_{\text{F}}^2}$, $\frac{\text{dist}_{2,\infty}^2(\hat{\boldsymbol{U}},\boldsymbol{U}^\star)}{\|\boldsymbol{U}^\star\|_{2,\infty}^2}$, and $\frac{\|\hat{\boldsymbol{T}}-\boldsymbol{T}^\star\|_\infty^2}{\|\boldsymbol{T}^\star\|_\infty^2}$) versus SNR. Figure 1(d) illustrates that all three types of relative squared errors scale inversely proportional to the SNR (since the slope in the figure is roughly $-1$), which is consistent with our statistical guarantees.

## 2.4. Notation

Before proceeding, we gather a few notations that will be used throughout this paper. First of all, for any matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, we let $\|\boldsymbol{M}\|$ and $\|\boldsymbol{M}\|_{\text{F}}$ denote the operator norm (or the spectral norm) and the Frobenius norm of $\boldsymbol{M}$, respectively, and we let $\boldsymbol{M}_{i,:}$ and $\boldsymbol{M}_{:,i}$ denote the $i$ th row and $i$ th column, respectively. In addition, we let $\lambda_1(\boldsymbol{M}) \geq \lambda_2(\boldsymbol{M}) \geq \cdots \geq \lambda_d(\boldsymbol{M})$ denote the eigenvalues of $\boldsymbol{M}$ and $\sigma_1(\boldsymbol{M}) \geq \sigma_2(\boldsymbol{M}) \geq \cdots \geq \sigma_d(\boldsymbol{M})$ denote the singular values of $\boldsymbol{M}$.

For any tensor $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$, let $\boldsymbol{T}_{i,:,:} \in \mathbb{R}^{d\times d}$ denote the mode-1 $i$-slice with entries $(\boldsymbol{T}_{i,:,:})_{j,k} = T_{i,j,k}$; $\boldsymbol{T}_{:,i,:}$ and $\boldsymbol{T}_{:,:,i}$ are defined in a similar way. For any tensors $\boldsymbol{T}, \boldsymbol{R} \in \mathbb{R}^{d\times d\times d}$, the inner product is defined as $\langle \boldsymbol{T}, \boldsymbol{R}\rangle := \sum_{j,k,l} T_{j,k,l}R_{j,k,l}$. The Frobenius norm of $\boldsymbol{T}$ is defined as $\|\boldsymbol{T}\|_{\text{F}} := \sqrt{\langle \boldsymbol{T}, \boldsymbol{T}\rangle}$. For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we define

the vector products of a tensor $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$—denoted by $\boldsymbol{T}\times_3 \boldsymbol{u} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{T}\times_1 \boldsymbol{u}\times_2 \boldsymbol{v} \in \mathbb{R}^{d}$—such that

$$[\boldsymbol{T}\times_3 \boldsymbol{u}]_{ij} := \sum\nolimits_{1\le k\le d} T_{i,j,k}u_k, \quad 1\le i,j\le d; \tag{20a}$$

$$[\boldsymbol{T}\times_1 \boldsymbol{u}\times_2 \boldsymbol{v}]_{k} := \sum\nolimits_{1\le i,j\le d} T_{i,j,k}u_i v_j, \quad 1\le k\le d. \tag{20b}$$

The products $\boldsymbol{T}\times_2 \boldsymbol{u} \in \mathbb{R}^{d\times d}$, $\boldsymbol{T}\times_3 \boldsymbol{u} \in \mathbb{R}^{d\times d}$, $\boldsymbol{T}\times_1 \boldsymbol{u}\times_3 \boldsymbol{v} \in \mathbb{R}^{d}$, and $\boldsymbol{T}\times_2 \boldsymbol{u}\times_3 \boldsymbol{v} \in \mathbb{R}^{d}$ are defined in a similar manner. For any $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{d\times r}$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r] \in \mathbb{R}^{d\times r}$, we further define

$$\boldsymbol{T} \times_1^{\text{seq}} \boldsymbol{U} \times_2^{\text{seq}} \boldsymbol{V} := [\boldsymbol{T}\times_1 \boldsymbol{u}_i \times_2 \boldsymbol{v}_i]_{1\le i\le r} \in \mathbb{R}^{d\times r}. \tag{21}$$

In addition, the operator norm of $\boldsymbol{T}$ is defined as

$$\|\boldsymbol{T}\| := \sup_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in\mathbb{S}^{d-1}} \langle \boldsymbol{T}, \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}\rangle, \tag{22}$$

where $\mathbb{S}^{d-1} := \{\boldsymbol{u}\in\mathbb{R}^d \mid \|\boldsymbol{u}\|_2 = 1\}$ indicates the unit sphere in $\mathbb{R}^d$.

Further, $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means that $|f(n)/g(n)| \le C_1$ for some constant $C_1 > 0$; $f(n) \gtrsim g(n)$ means that $|f(n)/g(n)| \ge C_2$ for some constant $C_2 > 0$; $f(n) \asymp g(n)$ means that $C_1 \le |f(n)/g(n)| \le C_2$ for some constants $C_1, C_2 > 0$; and $f(n) = o(g(n))$ means that $\lim_{n\to\infty} f(n)/g(n) = 0$. In addition, $f(n) \ll g(n)$ means that $f(n) \le c_1 g(n)$ for some sufficiently small constant $c_1 > 0$, and $f(n) \gg g(n)$ means that $f(n) \ge c_2 g(n)$ for some sufficiently large constant $c_2 > 0$.

## 3. Initialization

This section presents formal details of the proposed two-step initialization, accompanied by some intuition. We defer the discussion about alternative approaches to the supplementray materials. Recall that the proposed initialization procedure consists of two steps, which we discuss separately.

### 3.1. Step 1: Subspace Estimation via a Spectral Method

The spectral algorithm is often applied in conjunction with simple "unfolding" (or "matricization") to estimate the *subspace* spanned by the $r$ factors $\{\boldsymbol{u}_i^\star\}_{1\le i\le r}$. This strategy is partly motivated by prior approaches developed for covariance estimation with missing data (Lounici 2014, Montanari and Sun 2018, Cai et al. 2021). We next provide a brief introduction.

Let

$$\boldsymbol{A} = \text{unfold}^{1\times 2}\left(\frac{1}{p}\boldsymbol{T}\right) \in \mathbb{R}^{d\times d^2}, \quad \text{or more concisely} \quad \boldsymbol{A} = \text{unfold}\left(\frac{1}{p}\boldsymbol{T}\right) \in \mathbb{R}^{d\times d^2} \tag{23}$$

be the mode-1 matricization of $p^{-1}\boldsymbol{T}$ (i.e., $\frac{1}{p}T_{i,j,k} = A_{i,(j-1)d+k}$ for any $1\le i,j,k\le d$) (Kolda and Bader 2009). The rationale of this step is that, under our model, the unfolded matrix $\boldsymbol{A}$ obeys

$$\mathbb{E}[\boldsymbol{A}] = \text{unfold}(\boldsymbol{T}^\star) = \sum_{i=1}^{r} \boldsymbol{u}_i^\star (\boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star)^\top =: \boldsymbol{A}^\star, \tag{24}$$

whose column space is precisely the span of $\{\boldsymbol{u}^\star\}_{1\le i\le r}$. This motivates one to estimate the $r$-dimensional column space of $\mathbb{E}[\boldsymbol{A}]$ from $\boldsymbol{A}$. Toward this, a natural strategy is to look at the principal subspace of $\boldsymbol{A}\boldsymbol{A}^\top$. However, the diagonal entries of $\boldsymbol{A}\boldsymbol{A}^\top$ bear too much influence on the principal directions and need to be properly down-weighed. The current paper chooses to work with the principal subspace of the following matrix that zeros out all diagonal components:

$$\boldsymbol{B} := \mathcal{P}_{\text{off-diag}}(\boldsymbol{A}\boldsymbol{A}^\top), \tag{25}$$

where $\mathcal{P}_{\text{off-diag}}(\boldsymbol{Z})$ extracts out the off-diagonal entries of a squared matrix $\boldsymbol{Z}$. If we let $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ be an orthonormal matrix whose columns are the top-$r$ eigenvectors of $\boldsymbol{B}$, then $\boldsymbol{U}$ serves as our subspace estimate. See Algorithm 2 for a summary of the procedure.

### 3.2. Step 2: Retrieval of Low-Rank Tensor Factors from the Subspace Estimate

**3.2.1. Procedure.** As it turns out, it is possible to obtain rough (but reasonable) estimates of all individual low-rank tensor factors $\{\boldsymbol{u}_i^\star\}_{1\le i\le r}$—up to global permutation—given a reliable subspace estimate $\boldsymbol{U}$. This is in stark contrast to the low-rank matrix recovery case, where there exists some global rotational ambiguity that prevents us from disentangling the $r$ factors of interest.

We begin by describing how to retrieve *one* tensor factor from the subspace estimate—a procedure summarized in RETRIEVE-ONE-TENSOR-FACTOR(). Let us generate a random vector from the provided subspace $\boldsymbol{U}$ (which has orthonormal columns), that is,

$$\boldsymbol{\theta} = \underbrace{\boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{g}}_{\text{projection of } \boldsymbol{g} \text{ onto } \boldsymbol{U}}, \qquad \boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d). \tag{26}$$

The rescaled tensor data $p^{-1}\boldsymbol{T}$ is then transformed into a matrix via proper "projection" along this random direction $\boldsymbol{\theta}$, namely,

$$\boldsymbol{M} = \frac{1}{p}\boldsymbol{T}\times_3 \boldsymbol{\theta} \in \mathbb{R}^{d\times d}. \tag{27}$$

Our estimate for a tensor factor is then given by $\lambda^{1/3}\boldsymbol{\nu}$, where $\boldsymbol{\nu}$ is the leading singular vector of $\boldsymbol{M}$ obeying $\langle \boldsymbol{T}, \boldsymbol{\nu}^{\otimes 3}\rangle \ge 0$, and $\lambda$ is taken as $\lambda = \langle p^{-1}\boldsymbol{T}, \boldsymbol{\nu}^{\otimes 3}\rangle$. Informally, $\boldsymbol{\nu}$ reflects the direction of the component $\boldsymbol{u}_i^\star$ that exhibits the largest correlation with the random direction $\boldsymbol{\theta}$, and $\lambda$ forms an estimate of the corresponding size $\|\boldsymbol{u}_i^\star\|_2$.

A challenge remains, however, as there are oftentimes more than one tensor factor to estimate. To address this issue, we propose to rerun the aforementioned procedure multiple times, so as to ensure that we get to retrieve each tensor factor of interest at least

once. We will then apply a careful pruning procedure (i.e., PRUNE()) to remove redundancy.

**3.2.2. Intuition.** To develop some intuition about the aforementioned procedure, consider the "heuristic" case where $\boldsymbol{\theta} = \boldsymbol{U}^\star(\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star)^{-1}\boldsymbol{U}^{\star\top}\boldsymbol{g}$, namely, the idealistic scenario where the subspace estimate $\boldsymbol{U}$ is accurate. Averaging out the randomness in the sampling pattern and the noise, we see that the expected projected matrix (27) takes the following form:

$$\mathbb{E}[\boldsymbol{M}\,|\,\boldsymbol{\theta}] = \boldsymbol{T}^\star \times_3 \boldsymbol{\theta} = \sum_{i=1}^{r} \langle \boldsymbol{\theta}, \boldsymbol{u}_i^\star \rangle \boldsymbol{u}_i^\star \boldsymbol{u}_i^{\star\top}.$$

As a result, in the incoherent case where $\{\boldsymbol{u}_j^\star\}$ are nearly orthogonal to each other, the leading singular vector of $\mathbb{E}[\boldsymbol{M}|\boldsymbol{\theta}]$—and hence that of $\boldsymbol{M}$ (i.e., $\boldsymbol{w}$)—is expected to be reasonably close to the factor $\boldsymbol{u}_i^\star$ that enjoys the largest projected coefficient. In other words, we expect

$$\boldsymbol{v} \approx \frac{1}{\|\boldsymbol{u}_i^\star\|_2}\boldsymbol{u}_i^\star, \quad \text{where } i = \arg\max_{1\le j\le r} |\langle \boldsymbol{\theta}, \boldsymbol{u}_j^\star \rangle|. \tag{28}$$

In the mean time, armed with (28) and the incoherence assumption (such that $\boldsymbol{u}_i^\star$ and $\boldsymbol{u}_j^\star$ are nearly orthogonal for $i \neq j$), one might have

$$\begin{aligned} \lambda = \langle \boldsymbol{T}^\star, \boldsymbol{v}^{\otimes 3} \rangle &\approx \frac{1}{\|\boldsymbol{u}_i^\star\|_2^3} \langle \boldsymbol{T}^\star, \boldsymbol{u}_i^{\star\otimes 3} \rangle \\ &\approx \frac{1}{\|\boldsymbol{u}_i^\star\|_2^3} \langle \boldsymbol{u}_i^{\star\otimes 3}, \boldsymbol{u}_i^{\star\otimes 3} \rangle = \|\boldsymbol{u}_i^\star\|_2^3, \end{aligned} \tag{29}$$

thus explaining our choice of $\lambda$ in the proposed procedure. These arguments hint at the ability of our procedure in retrieving one tensor factor in each round.

This intuitive argument, however, does not explain why we need to first project a random vector $\boldsymbol{g}$ onto the (approximate) column space of $\boldsymbol{U}^\star$. Although we will not go into detailed calculations here, we remark in passing a crucial high variability issue: without proper projection, the perturbation incurred by both the missing data and the noise might far exceed the strength of the true signal. As a result, it is advised to first project the data onto the desired subspace, in the hope of amplifying the signal-to-noise ratio.

## 4. Related Work

One of the most natural ideas for solving tensor completion is to first unfold the tensor data into matrices, followed by proper convex relaxation commonly adopted for low-rank matrix completion. Given that there is more than one way to matricize a tensor, several prior works have explored the design of matrix norms that can exploit the tensor structure more effectively (Tomioka et al. 2010, Gandy et al. 2011, Liu et al. 2013, Romera-Paredes and Pontil 2013, Mu et al. 2014, Lu et al. 2016). Such algorithms have been robustified to enable reliable recovery against sparse outliers as well (Goldfarb and Qin 2014). For the most part, however, such unfolding-based convex relaxation necessarily incurs loss of structural information, which is particularly severe when handling odd-order tensors. The sample complexity developed for this paradigm is often suboptimal vis-à-vis the computational limits (i.e., minimal sample complexity achievable by polynomial-time algorithms).

Motivated by the aforementioned suboptimality issue, Yuan and Zhang (2016, 2017) proposed to minimize instead the tensor nuclear norm subject to data constraints, which provably allows for reduced sample complexity. The issue, however, is that computing the tensor nuclear norm itself is already computationally intractable, thus limiting its applicability to even moderate-dimensional problems. Similar findings have also been discovered for tensor atomic norm minimization (Driggs et al. 2019). When restricted to polynomial-time algorithms, the best statistical guarantees are often attained via convex relaxation tailored to the sum-of-squares hierarchy (Barak and Moitra 2016); the resulting computational cost, however, remains prohibitively high for practical large-scale problems. Another matrix nuclear norm minimization algorithm has been proposed based on promoting certain structures on certain factor matrices (Liu et al. 2014). Developing statistical guarantees is, however, not the focal point of this work.

Moving beyond convex relaxation, a number of prior papers have developed nonconvex algorithms for tensor completion, examples including iterative hard thresholding (Rauhut et al. 2017), alternating minimization (Jain and Oh 2014, Xu et al. 2015, Wang et al. 2016), tensor SVD (Zhang and Aeron 2017), optimization on manifold (Kasai and Mishra 2016, Steinlechner 2016, Xia and Yuan 2017), proximal average algorithm with nonconvex regularizer (Yao 2018), and block coordinate decent (Xu and Yin 2013, Ji et al. 2016). When it comes to the model considered herein, these algorithms either lack optimal statistical guarantees or come with a computational cost that is significantly higher than a linear-time algorithm.

The algorithm and theory that we develop are largely inspired by the recent advances of nonconvex optimization algorithms for low-rank matrix recovery problems (Keshavan et al. 2010a,b; Candès et al. 2015; Chen and Wainwright 2015; Sun and Luo 2016; Yi et al. 2016; Chen and Candès 2017). The main theoretical tool—the leave-one-out analysis—is a powerful technique that has proved successful in various other statistical problems (El Karoui 2015, Abbe et al. 2017, Ding and Chen 2018, Zhong and Boumal 2018, Chen et al. 2019c, Chen et al. 2019d, Chen et al. 2019e, Li et al. 2019, Pananjady and Wainwright 2019, Ma et al. 2020, Chen et al. 2021). There are several major differences

between the analysis of nonconvex tensor completion and that of nonconvex matrix recovery. For instance, our initialization scheme is substantially more complicated than the matrix recovery counterpart, thus requiring much more sophisticated analysis; in addition, the local convergence stage of tensor completion does not suffer from rotational ambiguity (which often appears in nonconvex matrix completion), and hence we only need to handle permutational ambiguity.

In addition, the current paper focuses on nonadaptive uniform random sampling. If there is freedom in designing the sampling mechanism, then one can often expect improved performance (see, e.g., Krishnamurthy and Singh 2013, Zhang 2019). Fundamental criteria that enable perfect low-CP-rank tensor completion have been studied by Ashraphijuo and Wang (2017).

Tensor completion is simply a special example of the tensor recovery literature. There is a large body of results tackling various other tensor recovery and estimation problems, including, but not limited to, tensor decomposition (Kolda 2001; Kolda and Bader 2009; Kim et al. 2013; Anandkumar et al. 2014a, b; Ge et al. 2015; Tang and Shah 2015; Hopkins et al. 2016; Ge and Ma 2017; Sidiropoulos et al. 2017; Sun et al. 2017; Zoubir et al. 2018), tensor SVD and factorization (Kilmer et al. 2013, Zhang and Aeron 2017, Zhang and Xia 2018), and tensor regression and sketching (Rauhut et al. 2017, Chen et al. 2019b, Hao et al. 2019, Hao et al. 2020). The algorithmic ideas explored in this paper might have implications for these tensor-related problems as well.

Finally, we remark that, compared with the conference version (Cai et al. 2019), the current paper (1) extends the results presented therein to a more general case, where both the rank $r$ and the incoherence parameter $\mu$ are allowed to grow with $d$; (2) discusses how to handle asymmetric tensors; and (3) explains in detail the inadequacy of other initialization schemes (including both random initialization and tensor power methods). More numerical experiments have also been carried out and reported.

## 5. Analysis

In this section, we outline the proof of Theorem 2 and defer the detailed proof to the e-companion. The analysis is divided into three parts:

- In Section 5.1, we show that, given an initial estimate sufficiently close to the ground truth, vanilla gradient descent converges linearly. These are formalized in Lemmas 3 and 6.
- Sections 5.2–5.3 provide statistical guarantees for the two steps of the initialization procedure (see Theorem 3).
- Under the assumptions of Theorem 2, one can see that the initialization satisfies the requirement of linear convergence of vanilla gradient descent. Therefore, Theorem 2 immediately follows from the results in Sections 5.1–5.3.

### 5.1. Analysis for Local Convergence of GD

In this section, we demonstrate that if the initialization is reasonably good, then vanilla gradient descent converges linearly to a solution with the desired statistical accuracy. We postpone the analysis for initialization to Sections 5.2–5.3 for convenience of presentation.

**5.1.1. Preliminaries: Gradient and Hessian Calculation.** First of all, using our notation $\times^{\text{seq}}$ defined in (21), we can write

$$\nabla f(\boldsymbol{U}) = \frac{1}{p}\mathcal{P}_{\Omega}\left(\sum_{1\le i\le r}\boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^{\star} - \boldsymbol{E}\right)\times_1^{\text{seq}}\boldsymbol{U}\times_2^{\text{seq}}\boldsymbol{U}. \tag{30}$$

Next, we find it convenient to define an auxiliary loss function $f_{\text{clean}}(\boldsymbol{U}):\mathbb{R}^{d\times r}\to\mathbb{R}_+$ that corresponds to the noiseless case:

$$f_{\text{clean}}(\boldsymbol{U}) = \frac{1}{6p}\left\|\mathcal{P}_{\Omega}\left(\sum_{1\le i\le r}\boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^{\star}\right)\right\|_{\text{F}}^2. \tag{31}$$

The gradient of $f_{\text{clean}}$ with respect to (w.r.t.) $\boldsymbol{u}_s$ $(1\le s\le r)$ is thus given by

$$\nabla_{\boldsymbol{u}_s} f_{\text{clean}}(\boldsymbol{U}) = \frac{1}{p}\mathcal{P}_{\Omega}\left(\sum_{1\le i\le r}\boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^{\star}\right)\times_1\boldsymbol{u}_s\times_2\boldsymbol{u}_s,\quad 1\le s\le r, \tag{32}$$

and hence one can write

$$\nabla f_{\text{clean}}(\boldsymbol{U}) = \frac{1}{p}\mathcal{P}_{\Omega}\left(\sum_{1\le i\le r}\boldsymbol{u}_i^{\otimes 3} - \boldsymbol{T}^{\star}\right)\times_1^{\text{seq}}\boldsymbol{U}\times_2^{\text{seq}}\boldsymbol{U}. \tag{33}$$

This clearly satisfies

$$\nabla f(\boldsymbol{U}) = \nabla f_{\text{clean}}(\boldsymbol{U}) - \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{E})\times_1^{\text{seq}}\boldsymbol{U}\times_2^{\text{seq}}\boldsymbol{U}. \tag{34}$$

Moreover, direct algebraic manipulations give that, for any matrix $\boldsymbol{V}=[\boldsymbol{v}_1,\ldots,\boldsymbol{v}_r]\in\mathbb{R}^{d\times r}$,

$$\begin{aligned}&\text{vec}(\boldsymbol{V})^{\top}\nabla^2 f_{\text{clean}}(\boldsymbol{U})\text{vec}(\boldsymbol{V})\\ &= \frac{1}{3p}\left\|\mathcal{P}_{\Omega}\left(\sum_{1\le s\le r}\boldsymbol{u}_s\otimes\boldsymbol{u}_s\otimes\boldsymbol{v}_s + \boldsymbol{u}_s\otimes\boldsymbol{v}_s\otimes\boldsymbol{u}_s\right.\right.\\ &\quad\left.\left. + \boldsymbol{v}_s\otimes\boldsymbol{u}_s\otimes\boldsymbol{u}_s\right)\right\|_{\text{F}}^2 + \frac{2}{p}\left\langle\mathcal{P}_{\Omega}\left(\sum_{s\in[r]}\boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^{\star}\right),\right.\\ &\quad\left.\times\sum_{s\in[r]}\boldsymbol{v}_s\otimes\boldsymbol{v}_s\otimes\boldsymbol{u}_s\right\rangle,\end{aligned} \tag{35}$$

where $\text{vec}(\boldsymbol{V})$ denotes the vectorization of $\boldsymbol{V}$.

**5.1.2. Local Strong Convexity and Smoothness.** At the heart of our analysis is a crucial geometric property of the objective function; that is, the noiseless loss function $f_{\text{clean}}$ behaves like a locally strongly convex and smooth function. This fact, which is formally

stated in the following lemma, is the key enabler of fast local convergence of vanilla GD.

**Lemma 1** (Local Strong Convexity and Smoothness).

$$p \geq c_0 \max\left\{\frac{\log^3 d}{d^{3/2}}, \frac{\mu^2 r^2 \log d}{d^2}\right\} \qquad r \leq c_1 \sqrt{\frac{d}{\mu}} \tag{36}$$

*for some sufficiently large (respectively, small) constant $c_0 > 0$ (respectively, $c_1 > 0$). Then with probability greater than $1 - O(d^{-10})$,*

$$\frac{1}{2}\lambda_{\min}^{\star 4/3} \| \boldsymbol{V} \|_{\mathrm{F}}^2 \leq \mathsf{vec}(\boldsymbol{V})^\top \nabla^2 f_{\mathsf{clean}}(\boldsymbol{U})\, \mathsf{vec}(\boldsymbol{V}) \leq 4\lambda_{\max}^{\star 4/3} \| \boldsymbol{V} \|_{\mathrm{F}}^2 \tag{37}$$

*holds simultaneously for all $\boldsymbol{V} \in \mathbb{R}^{d\times r}$ and all $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ obeying*

$$\|\boldsymbol{U} - \boldsymbol{U}^\star\|_{\mathrm{F}} \leq \delta \| \boldsymbol{U}^\star \|_{\mathrm{F}} \quad \text{and} \quad \| \boldsymbol{U} - \boldsymbol{U}^\star \|_{2,\infty} \leq \delta \| \boldsymbol{U}^\star \|_{2,\infty}. \tag{38}$$

*Here, $\delta \leq c_2/(\mu^{3/2} r)$ for some sufficiently small constant $c_2 > 0$.*

In order to invoke Lemma 1, one needs to make sure that the decision matrix $\boldsymbol{U}$ of interest (e.g., $\boldsymbol{U}^t$ in the GD sequence) satisfies the condition (38). This, however, is a fairly stringent condition, as it requires $\boldsymbol{U}$ to be close to the truth in every single row.

**5.1.3. Leave-One-Out Gradient Descent Sequences.** Motivated by the analytical framework developed for low-rank matrix recovery (Ma et al. 2017, Chen et al. 2019a), we introduce the following leave-one-out sequences, which play a crucial role in guaranteeing that the entire trajectory $\{\boldsymbol{U}^t\}_{t\geq 0}$ satisfies the condition (38), as required in Lemma 1.

Specifically, we define for each $1 \leq m \leq d$ the following auxiliary loss function:

$$\begin{aligned} f^{(m)}(\boldsymbol{U}) \triangleq \frac{1}{6p}\left\| \mathcal{P}_{\Omega_{-m}}\Big(\sum_{1\leq s\leq r} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^\star - \boldsymbol{E}\Big)\right\|_{\mathrm{F}}^2 \\ + \frac{1}{6}\left\| \mathcal{P}_{m}\Big(\sum_{1\leq s\leq r} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^\star\Big)\right\|_{\mathrm{F}}^2, \end{aligned} \tag{39}$$

where

- $\mathcal{P}_{\Omega_m}$ is the projection onto the subspace of tensors supported on $\{(i,j,k) \in \Omega : i = m \text{ or } j = m \text{ or } k = m\}$;
- $\mathcal{P}_{\Omega_{-m}}$ is the projection onto the subspace of tensors supported on $\{(i,j,k) \in \Omega : i \neq m \text{ and } j \neq m \text{ and } k \neq m\}$;
- $\mathcal{P}_m$ is the projection onto the subspace of tensors supported on $\{(i,j,k) \in [d]^3 : i = m \text{ or } j = m \text{ or } k = m\}$.

In words, this function is obtained by replacing all data at locations $\{(i,j,k) \in [d]^3 : i = m \text{ or } j = m \text{ or } k = m\}$ by their expected values, thus removing all randomness associated with this location subset. The gradient of $f^{(m)}(\boldsymbol{U})$ w.r.t. $\boldsymbol{u}_s$ $(1 \leq s \leq r)$ can be computed as

$$\begin{aligned} \nabla_{\boldsymbol{u}_s} f^{(m)}(\boldsymbol{U}) = \frac{1}{p}\mathcal{P}_{\Omega_{-m}}\Big(\sum_{1\leq s\leq r} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^\star - \boldsymbol{E}\Big) \\ \times_1 \boldsymbol{u}_s \times_2 \boldsymbol{u}_s + \mathcal{P}_m\Big(\sum_{1\leq s\leq r} \boldsymbol{u}_s^{\otimes 3} - \boldsymbol{T}^\star\Big) \\ \times_1 \boldsymbol{u}_s \times_2 \boldsymbol{u}_s. \end{aligned} \tag{40}$$

We then denote by $\{\boldsymbol{U}^{t,(m)}\}_{t\geq 0}$ the iterative sequence obtained by running gradient descent w.r.t. the leave-one-out loss $f^{(m)}(\cdot)$ (see Algorithm 4). By construction, as long as $\boldsymbol{U}^{0,(m)}$ is independent of the sampling locations and the noise associated with the locations $\{(i,j,k) \in \Omega : i = m \text{ or } j = m \text{ or } k = m\}$ (which holds true as detailed momentarily), then the entire trajectory $\{\boldsymbol{U}^{t,(m)}\}_{t\geq 0}$ becomes statistically independent of such randomness. This is a crucial property that allows us to decouple the complicated statistical dependency.

**Algorithm 4** (The $m$ th Leave-One-Out Sequence)

1: Generate an initial estimate $\boldsymbol{U}^{0,(m)}$ via Algorithm 5.

2: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

3: $\boldsymbol{U}^{t+1,(m)} = \boldsymbol{U}^{t,(m)} - \eta_t \nabla f^{(m)}(\boldsymbol{U}^{t,(m)})$.

**5.1.4. Key Lemmas.** The proof for local linear convergence of GD is inductive in nature, which proceeds on the basis of the following set of inductive hypotheses. As we shall see in Corollary 3 in Section 5.3, this set of inductive hypotheses—modulo some global permutation—is valid with high probability when $t = 0$. In order to simplify presentation, we remove the consideration of the global permutation factor throughout this section (i.e., we assume that the following holds for $\boldsymbol{U}^0\boldsymbol{\Pi}^0$ with some permutation matrix $\boldsymbol{\Pi}^0 \in \mathbb{R}^{r\times r}$ obeying $\boldsymbol{\Pi}^0 = \boldsymbol{I}$. Our key inductive hypotheses for the gradient update stage are summarized as follows:

$$\|\boldsymbol{U}^t - \boldsymbol{U}^\star\|_{\mathrm{F}} \leq \left(C_1\rho^t\mathcal{E}_{\mathsf{local}} + C_2\frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\right)\|\boldsymbol{U}^\star\|_{\mathrm{F}}; \tag{41a}$$

$$\|\boldsymbol{U}^t - \boldsymbol{U}^\star\|_{2,\infty} \leq \left(C_3\rho^t\mathcal{E}_{\mathsf{local}} + C_4\frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\right) \times \|\boldsymbol{U}^\star\|_{2,\infty}; \tag{41b}$$

$$\|\boldsymbol{U}^t - \boldsymbol{U}^{t,(m)}\|_{\mathrm{F}} \leq \left(C_5\rho^t\mathcal{E}_{\mathsf{local}} + C_6\frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\right) \times \|\boldsymbol{U}^\star\|_{2,\infty}; \tag{41c}$$

$$\left\|\left(\boldsymbol{U}^{t,(m)} - \boldsymbol{U}^\star\right)_{m,:}\right\|_2 \leq \left(C_7\rho^t\mathcal{E}_{\mathsf{local}} + C_8\frac{\sigma}{\lambda_{\min}^\star}\sqrt{\frac{d\log d}{p}}\right) \times \|\boldsymbol{U}^\star\|_{2,\infty}; \tag{41d}$$

for some quantity $\mathcal{E}_{\mathsf{local}} > 0$ (depending possibly on $\mu$ and $r$) and some constants $C_1, \cdots, C_8 > 0$. There exist a few straightforward consequences of the hypotheses (41), which we record in the following lemma.

**Lemma 2.** *Assume that the hypotheses (41) hold. Then we have*

$$\|\boldsymbol{U}^{t,(m)}-\boldsymbol{U}^{\star}\|_{\mathrm{F}}\le\left(2C_1\rho^t\mathcal{E}_{\mathrm{local}}+2C_2\frac{\sigma}{\lambda^{\star}_{\min}}\sqrt{\frac{d\log d}{p}}\right)\times\|\boldsymbol{U}^{\star}\|_{\mathrm{F}}, \tag{42}$$

$$\|\boldsymbol{U}^{t,(m)}-\boldsymbol{U}^{\star}\|_{2,\infty}\le\left((C_3+C_5)\rho^t\mathcal{E}_{\mathrm{local}}+(C_4+C_6)\times\frac{\sigma}{\lambda^{\star}_{\min}}\sqrt{\frac{d\log d}{p}}\right)\|\boldsymbol{U}^{\star}\|_{2,\infty}. \tag{43}$$

Our proof for the hypotheses (41) is inductive in nature: we would like to show that if the hypotheses in (41) hold for the $t$ th iteration, then they continue to be valid for the $(t+1)$ th iteration. We shall justify each of the aforementioned hypotheses inductively through the following lemmas.

**Lemma 3.** *Suppose that*

$$p\ge c_0\frac{\mu^3r^2\log^3 d}{d^{3/2}},\quad \frac{\sigma}{\lambda^{\star}_{\min}}\le c_1\frac{\sqrt{p}}{\mu^{3/2}r\sqrt{d\log d}},\quad\text{and}\quad r\le c_2\sqrt{\frac{d}{\mu}}$$

*for some sufficiently large constant $c_0>0$ and some sufficiently small constant $c_1,c_2>0$. Assume that the hypotheses (41) hold for the tth iteration and $\mathcal{E}_{\mathrm{local}}\le c_3/(\mu^{3/2}r)$ for some sufficiently small constant $c_3>0$. Then with probability at least $1-O(d^{-10})$,*

$$\|\boldsymbol{U}^{t+1}-\boldsymbol{U}^{\star}\|_{\mathrm{F}}\le\left(C_1\rho^{t+1}\mathcal{E}_{\mathrm{local}}+C_2\frac{\sigma}{\lambda^{\star}_{\min}}\sqrt{\frac{d\log d}{p}}\right)\times\|\boldsymbol{U}^{\star}\|_{\mathrm{F}}, \tag{44}$$

*provided that $0<\eta\le\lambda^{\star 4/3}_{\min}/(32\lambda^{\star 8/3}_{\max})$, $1-(\lambda^{\star 4/3}_{\min}/5)\eta\le\rho<1$, and $C_2$ is sufficiently large.*

**Lemma 4.** *Suppose that*

$$p\ge c_0\frac{\mu^3r^2\log^3 d}{d^{3/2}},\quad \frac{\sigma}{\lambda^{\star}_{\min}}\le c_1\frac{\sqrt{p}}{\mu^{3/2}r\sqrt{d\log d}},\ r\le c_2\sqrt{\frac{d}{\mu}}$$

*for some sufficiently large constant $c_0>0$ and some sufficiently small constant $c_1,c_2>0$. Assume that the hypotheses (41) hold for the tth iteration and $\mathcal{E}_{\mathrm{local}}\le c_3/(\mu^{3/2}r)$ for some sufficiently small constant $c_3>0$. Then with probability at least $1-O(d^{-10})$, one has*

$$\|\boldsymbol{U}^{t+1,(m)}-\boldsymbol{U}^{t+1}\|_{\mathrm{F}}\le\left(C_5\rho^{t+1}\mathcal{E}_{\mathrm{local}}+C_6\frac{\sigma}{\lambda^{\star}_{\min}}\sqrt{\frac{d\log d}{p}}\right)\times\|\boldsymbol{U}^{\star}\|_{2,\infty}, \tag{45}$$

*provided that $0<\eta\le\lambda^{\star 4/3}_{\min}/(32\lambda^{\star 8/3}_{\max})$, $1-(\lambda^{\star 4/3}_{\min}/5)\eta\le\rho<1$, and $C_6$ is sufficiently large.*

**Lemma 5.** *Suppose that*

$$p\ge c_0\frac{\mu^3r^2\log^3 d}{d^{3/2}},\quad \frac{\sigma}{\lambda^{\star}_{\min}}\le c_1\frac{\sqrt{p}}{\mu^{3/2}r\sqrt{d\log d}},\ r\le c_2\sqrt{\frac{d}{\mu}}$$

*for some sufficiently large constant $c_0>0$ and some sufficiently small constant $c_1,c_2>0$. Assume that the hypotheses (41) hold for the t th iteration and $\mathcal{E}_{\mathrm{local}}\le c_3/(\mu^{3/2}r)$ for some sufficiently small constant $c_3>0$. Then with probability at least $1-O(d^{-10})$, one has*

$$\left\|\left(\boldsymbol{U}^{t+1,(m)}-\boldsymbol{U}^{\star}\right)_{m,:}\right\|_2\le\left(C_7\rho^{t+1}\mathcal{E}_{\mathrm{local}}+C_8\frac{\sigma}{\lambda^{\star}_{\min}}\sqrt{\frac{d\log d}{p}}\right)\|\boldsymbol{U}^{\star}\|_{2,\infty}, \tag{46}$$

*provided that $0<\eta\le\lambda^{\star 4/3}_{\min}/(32\lambda^{\star 8/3}_{\min})$, $1-(\lambda^{\star 4/3}_{\min}/5)\eta\le\rho<1$, and $C_7$ and $C_8$ are sufficiently large.*

**Lemma 6.** *Suppose that*

$$p\ge c_0\frac{\mu^3r^2\log^3 d}{d^{3/2}},\quad \frac{\sigma}{\lambda^{\star}_{\min}}\le c_1\frac{\sqrt{p}}{\mu^{3/2}r\sqrt{d\log d}},\quad\text{and}\quad r\le c_2\sqrt{\frac{d}{\mu}}$$

*for some sufficiently large constant $c_0>0$ and some sufficiently small constant $c_1,c_2>0$. Assume that the hypotheses (41) hold for the tth iteration and $\mathcal{E}_{\mathrm{local}}\le c_3/(\mu^{3/2}r)$ for some sufficiently small constant $c_3>0$. Then with probability at least $1-O(d^{-10})$, one has*

$$\|\boldsymbol{U}^{t+1}-\boldsymbol{U}^{\star}\|_{2,\infty}\le\left(C_3\rho^{t+1}\mathcal{E}_{\mathrm{local}}+C_4\frac{\sigma}{\lambda^{\star}_{\min}}\sqrt{\frac{d\log d}{p}}\right)\|\boldsymbol{U}^{\star}\|_{2,\infty}, \tag{47}$$

*provided that $0<\eta\le\lambda^{\star 4/3}_{\min}/(32\lambda^{\star 8/3}_{\min})$, $1-(\lambda^{\star 4/3}_{\min}/5)\eta\le\rho<1$, and $C_3/(C_5+C_7)$ and $C_4/(C_6+C_8)$ are both sufficiently large.*

### 5.2. Analysis for Initialization: Part 1 (Subspace Estimation)

**5.2.1. Key Results.** The aim of this subsection is to demonstrate that the subspace estimate $\boldsymbol{U}$ computed by Algorithm 2 is sufficiently close to the space spanned by the true tensor factors. Given that the columns of $\boldsymbol{U}^{\star}=[\boldsymbol{u}^{\star}_1,\dots,\boldsymbol{u}^{\star}_r]$ are in general not orthogonal to each other, we shall define $\boldsymbol{U}^{\star}_{\mathrm{orth}}\in\mathbb{R}^{d\times r}$ as follows (obtained by proper orthonormalization):

$$\boldsymbol{U}^{\star}_{\mathrm{orth}}:=\boldsymbol{U}^{\star}(\boldsymbol{U}^{\star\top}\boldsymbol{U}^{\star})^{-\frac{1}{2}}. \tag{48}$$

This matrix $\boldsymbol{U}^{\star}_{\mathrm{orth}}$ reflects the rank-$r$ principal subspace of $\boldsymbol{A}^{\star}\boldsymbol{A}^{\star\top}=\sum_i\|\boldsymbol{u}^{\star}_i\|_2^4\boldsymbol{u}^{\star}_i\boldsymbol{u}^{\star\top}_i$, where we recall that $\boldsymbol{A}^{\star}\in\mathbb{R}^{d\times d^2}$ is the mode-1 matricization of $\boldsymbol{T}^{\star}$. In addition, we define the rotation matrix

$$\boldsymbol{R}:=\arg\min_{\boldsymbol{Q}\in\mathcal{O}^{r\times r}}\|\boldsymbol{U}\boldsymbol{Q}-\boldsymbol{U}^{\star}_{\mathrm{orth}}\|_{\mathrm{F}}, \tag{49}$$

where $\mathcal{O}^{r\times r}$ stands for the set of $r\times r$ orthonormal matrices. This can be viewed as the global rotation matrix that best aligns the two subspaces represented by $\boldsymbol{U}$ and $\boldsymbol{U}^\star_{\mathsf{orth}}$, respectively.

Equipped with the aforementioned notation, we can invoke (Cai et al. 2021, corollary 1) to arrive at the following lemma, which upper-bounds the distance between our subspace estimate $\boldsymbol{U}$ and the ground truth $\boldsymbol{U}^\star_{\mathsf{orth}}$.

**Lemma 7.** *There exist some universal constants $c_0, c_1, c_2 > 0$ such that if*

$$p \ge c_0 \frac{\mu^2 r \log^2 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda^\star_{\min}} \le c_1 \frac{\sqrt{p}}{d^{3/4}\sqrt{\log d}}, \quad \text{and} \quad r \le c_2 \sqrt{\frac{d}{\mu}},$$

*then with probability $1 - O(d^{-10})$, the subspace estimate $\boldsymbol{U}$ computed by* Algorithm 2 *obeys*

$$\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star_{\mathsf{orth}} \| \lesssim \mathcal{E}_{\mathsf{se}}, \tag{50a}$$

$$\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star_{\mathsf{orth}} \|_{2,\infty} \lesssim \mathcal{E}_{\mathsf{se}} \sqrt{\frac{\mu r}{d}}, \tag{50b}$$

*where $\boldsymbol{U}^\star_{\mathsf{orth}}$ and $\boldsymbol{R}$ are defined, respectively, in* (48) *and* (49), *and*

$$\mathcal{E}_{\mathsf{se}} := \frac{\mu^2 r \log d}{d^{3/2} p} + \sqrt{\frac{\mu^2 r \log d}{d^2 p}} + \frac{\sigma^2}{\lambda^{\star 2}_{\min}} \frac{d^{3/2} \log d}{p} + \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{d \log d}{p}} + \frac{\mu r}{d}. \tag{51}$$

In a nutshell, Lemma 7 asserts that, under our sample size, noise, and rank conditions, Algorithm 2 produces reliable estimates of the subspace spanned by the low-rank tensor factors $\{\boldsymbol{u}^\star_i\}_{1\le i\le r}$. The theorem quantifies the subspace distance in terms of both the spectral norm and $\|\cdot\|_{2,\infty}$, where the latter bound often reflects a considerably stronger sense of proximity compared with the former one.

As it turns out, in order to facilitate analysis for the subsequent stages, we need to introduce a certain leave-one-out sequences as well, which we detail in the next subsection.

**5.2.2. Leave-One-Out Sequences for Subspace Estimation.** The key idea of the leave-one-out analysis is to create auxiliary leave-one-out sequences that are (1) independent of a small fraction of the data and (2) sufficiently close to the true estimates. We introduce the following auxiliary tensor and $d\times d^2$-dimensional matrix for each $1 \le m \le d$:

$$\boldsymbol{T}^{(m)} := \mathcal{P}_{\Omega_{-m}}(\boldsymbol{T}) + p\mathcal{P}_m(\boldsymbol{T}^\star) \in \mathbb{R}^{d\times d\times d}, \tag{52}$$

$$\boldsymbol{A}^{(m)} := \text{mode-1 matricization of } \frac{1}{p}\boldsymbol{T}^{(m)}. \tag{53}$$

By construction, $\boldsymbol{T}^{(m)}$ and $\boldsymbol{A}^{(m)}$ are independent of $\mathcal{P}_{\Omega_m}(\boldsymbol{E})$, where we recall that

$$\Omega_{-m} := \{(i,j,k) \in \Omega : i \ne m \text{ and } j \ne m \text{ and } k \ne m\}, \tag{54}$$

$$\Omega_m := \{(i,j,k) \in \Omega : i = m \text{ or } j = m \text{ or } k = m\}. \tag{55}$$

We are now ready to introduce the auxiliary leave-one-out procedure for subspace estimation. Similar to the matrix $\boldsymbol{B}$ in Algorithm 2 (whose eigenspace serves as an estimate of the column space of $\boldsymbol{U}^\star$), we define an auxiliary matrix $\boldsymbol{B}^{(m)} \in \mathbb{R}^{d\times d}$ as follows:

$$\boldsymbol{B}^{(m)} = \mathcal{P}_{\mathsf{off\text{-}diag}}\left(\boldsymbol{A}^{(m)}\boldsymbol{A}^{(m)\top}\right), \tag{56}$$

where $\mathcal{P}_{\mathsf{off\text{-}diag}}(\cdot)$ (as already defined in Section 3.1) extracts out off-diagonal entries from a matrix. The rationale is simple: it can be easily verified that

$$\mathbb{E}[\boldsymbol{B}^{(m)}] = \boldsymbol{B}^\star - \mathcal{P}_{\mathsf{diag}}(\boldsymbol{B}^\star), \quad \boldsymbol{B}^\star := \boldsymbol{A}^\star\boldsymbol{A}^{\star\top}, \tag{57}$$

where $\mathcal{P}_{\mathsf{diag}}(\cdot)$ extracts out the diagonal entries of the matrix. This gives hope that the eigenspace of $\boldsymbol{B}^{(m)}$ is also a reliable estimate of the column space of $\boldsymbol{U}^\star$, provided that the diagonal entries of $\boldsymbol{B}^\star$ are sufficiently small. Consequently, we shall compute $\boldsymbol{U}^{0,(m)} \in \mathbb{R}^{d\times r}$—a matrix whose columns are the top-$r$ leading eigenvectors of $\boldsymbol{B}^{(m)}$. The procedure is summarized in Algorithm 5.

**Algorithm 5** (The $m$ th Leave-One-Out Sequence for Spectral Initialization)

1: Let $\boldsymbol{U}^{(m)}\boldsymbol{\Lambda}^{(m)}\boldsymbol{U}^{(m)\top}$ be the rank-$r$ eigen-decomposition of $\boldsymbol{B}^{(m)}$ defined in (56).

2: Generate the initial estimate $\boldsymbol{U}^{0,(m)} \in \mathbb{R}^{d\times r}$ from $\boldsymbol{U}^{(m)} \in \mathbb{R}^{d\times r}$ using Algorithm 6.

The following lemma plays a crucial role in our analysis, which formalizes the fact that the leave-one-out version $\boldsymbol{U}^{(m)}$ obtained by Algorithm 5 is extremely close to $\boldsymbol{U}$.

**Lemma 8.** *There exist some universal constants $c_0, c_1, c_2 > 0$ such that if*

$$p \ge c_0 \frac{\mu^2 r \log^2 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda^\star_{\min}} \le c_1 \frac{\sqrt{p}}{d^{3/4}\sqrt{\log d}}, \quad \text{and} \quad r \le c_2 \sqrt{\frac{d}{\mu}},$$

*then with probability $1 - O(d^{-10})$, the subspace estimate $\boldsymbol{U}^{(m)}$ computed by* Algorithm 5 *obeys*

$$\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\|_{\mathrm{F}} \lesssim \mathcal{E}_{\mathsf{loo}} \sqrt{\frac{\mu r}{d}} \tag{58}$$

*simultaneously for all $1 \le m \le d$, where*

$$\mathcal{E}_{\mathsf{loo}} := \frac{\mu^2 r \log d}{d^{3/2} p} + \sqrt{\frac{\mu^2 r \log d}{d^2 p}} + \frac{\sigma^2}{\lambda^{\star 2}_{\min}} \frac{d^{3/2} \log d}{p} + \frac{\sigma}{\lambda^\star_{\min}} \sqrt{\frac{d \log d}{p}}. \tag{59}$$

Lemma 8 follows immediately from the analysis of (Cai et al. 2021, lemma 4). As a remark, the construction of the leave-one-out sequences herein is slightly different from the one in (Cai et al. 2021). However, it is straightforward to adapt the proof of (Cai et al. 2021) to the case considered herein. We therefore omit the proof for the sake of brevity.

### 5.3. Analysis for Initialization: Part 2 (Retrieval of Individual Tensor Factors)

**5.3.1. Main Results and Leave-One-Out Sequences.** This section justifies that the procedure presented in Algorithm 3 allows to disentangle the tensor factors. For notational simplicity, we let

$$\overline{\boldsymbol{u}}_i^\star := \boldsymbol{u}_i^\star / \| \boldsymbol{u}_i^\star \|_2, \quad \lambda_i^\star := \| \boldsymbol{u}_i^\star \|_2^3, \quad 1 \le i \le d. \tag{60}$$

Our result is the following.

**Theorem 3.** *Fix any arbitrary small constant $\delta > 0$. Assume that*

$$\begin{aligned} p &\ge c_0 \frac{\mu^2 r^4 \log^4 d}{d^{3/2}}, \quad \frac{\sigma}{\lambda_{\min}^\star} \le c_1 \frac{\sqrt{p}}{r^{3/2} d^{3/4} \log^2 d}, \\ r &\le c_2 \left( \frac{d}{\mu^6 \log^6 d} \right)^{1/6}, \\ L &= c_3 r^{2\kappa^2} \log^{3/2} r, \\ \epsilon_{\mathrm{th}} &= c_4 \left\{ \frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{rd \log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}} \right\} \end{aligned} \tag{61}$$

*for some sufficiently large universal constant $c_0, c_3 > 0$ and some sufficiently small universal constants $c_1, c_2, c_4 > 0$. Then, with probability exceeding $1 - \delta$, there exists a permutation $\pi(\cdot) : [d] \mapsto [d]$ such that for all $1 \le i \le r$, the tensor factors $\{\boldsymbol{w}^i\}_{i=1}^r$ returned by* Algorithm 3 *satisfy*

$$\| \boldsymbol{w}^i - \overline{\boldsymbol{u}}_{\pi(i)}^\star \|_2 \lesssim \frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{rd \log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}}; \tag{62a}$$

$$\begin{aligned} &\| \boldsymbol{w}^i - \overline{\boldsymbol{u}}_{\pi(i)}^\star \|_\infty \\ &\lesssim \left\{ \frac{\mu^2 r \log^4 d}{d^{3/2} p} + \frac{\mu r \log^3 d}{d\sqrt{p}} + \frac{\sigma^2}{\lambda_{\min}^{\star 2}} \frac{d^{3/2} \log^4 d}{p} \right. \\ &\quad \left. + \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{rd \log^6 d}{p}} + \sqrt{\frac{\mu r \log^2 d}{d}} \right\} \sqrt{\frac{\mu r}{d}}; \end{aligned} \tag{62b}$$

$$|\lambda_i - \lambda_{\pi(i)}^\star| \lesssim \left\{ \frac{\mu r \log d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{rd \log^2 d}{p}} + \sqrt{\frac{\mu r \log d}{d}} \right\} \lambda_{\pi(i)}^\star. \tag{62c}$$

In short, this theorem asserts that the estimates returned by Algorithm 3 are—up to global permutation—reasonably close to the ground truth under our sample size and noise conditions. In order to establish this theorem, and in order to provide initial guesses for the leave-one-out GD sequences, we need to produce a leave-one-out sequence tailored to this part of the algorithm. Such auxiliary sequences are generated in a similar spirit as the previous ones, and we summarize them in Algorithm 6. As usual, the resulting leave-one-out estimates $\{\lambda_i^{(m)}, \boldsymbol{w}^{i,(m)}\}_{i=1}^r$ are statistically independent of $\mathcal{P}_{\Omega_m}(\boldsymbol{E})$.

In what follows, we gather a few key properties of the leave-one-out estimates, which play a crucial role in the analysis.

**Algorithm 6** (The $m$ th Leave-One-Out Sequence for Retrieving Individual Tensor Components)

1: **Input:** restart number $L$, threshold $\epsilon_{\mathrm{th}}$, subspace estimate $\boldsymbol{U}^{(m)}$ given by Algorithm 5.
2: **for** $\tau = 1, \ldots, L$ **do**
3: Recall the Gaussian vector $\boldsymbol{g}^\tau \sim \mathcal{N}(0, \boldsymbol{I}_d)$ generated in Algorithm 3.
4: $(\boldsymbol{\nu}^{\tau,(m)}, \lambda_\tau^{(m)}, \mathrm{spec\text{–}gap}_\tau^{(m)}) \leftarrow$ RETRIEVE-ONE-TENSOR-FACTOR$(\boldsymbol{T}^{(m)}, p, \boldsymbol{U}^{(m)}, \boldsymbol{g}^\tau)$.
5: Generate tensor factor estimates

$$\begin{aligned} &\{(\boldsymbol{w}^{1,(m)}, \lambda_1^{(m)}), \ldots, (\boldsymbol{w}^{r,(m)}, \lambda_r^{(m)})\} \\ &\quad \leftarrow \mathrm{PRUNE}(\{(\boldsymbol{\nu}^{\tau,(m)}, \lambda_\tau^{(m)}, \mathrm{spec\text{-}gap}_\tau^{(m)})\}_{\tau=1}^L, \epsilon_{\mathrm{th}}). \end{aligned}$$

6: **Output:** an initial estimate $\boldsymbol{U}^{0,(m)} = [(\lambda_1^{(m)})^{1/3} \boldsymbol{w}^{1,(m)}, \ldots, (\lambda_r^{(m)})^{1/3} \boldsymbol{w}^{r,(m)}]$.

**Theorem 4.** *Fix any arbitrarily small constant $\delta > 0$. Instate the assumptions in* Theorem 3. *With probability exceeding $1 - \delta$, the permutation function stated in Theorem 3 obeys that, for all $1 \le i \le r$ and all $1 \le m \le d$,*

$$\begin{aligned} \| \boldsymbol{w}^i - \boldsymbol{w}^{i,(m)} \|_2 &\lesssim \left\{ \frac{\mu^2 r \log^{3/2} d}{d^{3/2} p} + \frac{\mu \sqrt{r} \log d}{d\sqrt{p}} + \frac{\sigma^2}{\lambda_{\min}^{\star 2}} \frac{d^{3/2} \log^{3/2} d}{p} \right. \\ &\quad \left. + \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log^2 d}{p}} \right\} \sqrt{\frac{\mu r}{d}}; \end{aligned} \tag{63a}$$

$$\begin{aligned} |\lambda_i - \lambda_i^{(m)}| &\lesssim \left\{ \frac{\mu^2 r \log^{3/2} d}{d^{3/2} p} + \frac{\mu \sqrt{r} \log d}{d\sqrt{p}} + \frac{\sigma^2}{\lambda_{\min}^{\star 2}} \frac{d^{3/2} \log^{3/2} d}{p} \right. \\ &\quad \left. + \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{d \log^2 d}{p}} \right\} \sqrt{\frac{\mu r}{d}} \lambda_{\max}^\star; \end{aligned} \tag{63b}$$

$$\begin{aligned} \left| \left( \boldsymbol{w}^{i,(m)} - \overline{\boldsymbol{u}}_{\pi(i)}^\star \right)_m \right| &\lesssim \left\{ \frac{\sqrt{\mu r} \log^{7/2} d}{d^{3/2} p} + \frac{\mu r \log^3 d}{d\sqrt{p}} + \frac{\sigma}{\lambda_{\min}^\star} \frac{\log^4 d}{p} \right. \\ &\quad \left. + \frac{\sigma}{\lambda_{\min}^\star} \sqrt{\frac{rd \log^6 d}{p}} + \sqrt{\frac{\mu r \log^2 d}{d}} \right\} \sqrt{\frac{\mu r}{d}}. \end{aligned} \tag{63c}$$

With Theorems 3–4 in place, we can immediately establish a few desired properties (particularly those specified in Section 5.1) of our initial estimate, as asserted in the following corollary.

**Corollary 3.** *Fix any arbitrarily small constant $\delta > 0$. Instate the assumptions in* Theorem 2. *With probability exceeding $1 - \delta$, the estimates $\boldsymbol{U}^0$ and $\boldsymbol{U}^{0,(m)}$ returned by* Algorithm 3 *and* Algorithm 6, *respectively, satisfy the hypotheses* (41) *for $t = 0$.*

## 6. Discussion

The current paper uncovers the possibility of efficiently and stably completing a low-CP-rank tensor from partial and noisy entries. Perhaps somewhat unexpectedly, despite the high degree of nonconvexity, this problem can be solved to optimal statistical accuracy within nearly linear time, provided that the tensor of interest is well conditioned, incoherent, and of constant rank. To the best of our knowledge, this intriguing message has not been shown in the prior literature.

Moving forward, one pressing issue is to understand how to improve the algorithmic and theoretical dependency upon the tensor rank $r$ of the proposed method. Ideally, one would desire a fast algorithm whose sample complexity scales as $rd^{1.5}$, an order that is provably achievable by the sum-of-squares hierarchy. Additionally, in contrast to the matrix counterpart where the rank is upper bounded by the matrix dimension, the tensor CP rank is allowed to rise above $d$, which is commonly referred to as the *over-complete case*. Unfortunately, our current initialization scheme (i.e., the spectral method) fails to work unless $r < d$, and our local analysis for GD falls of accommodating the scenario with $r > d$. It would be of great interest to develop more powerful algorithms—in addition to more refined analysis—to tackle such an important over-complete regime.

Another tantalizing research direction is the exploration of landscape design for tensor completion. As our heuristic discussions as well as other prior work (e.g., Richard and Montanari 2014) suggest, randomly initialized gradient descent tailored to (4) seems unlikely to work, unless the sample size is significantly larger than the computational limit. This might mean either that there exist spurious local minima in the natural nonconvex least-squares formulation (4), or that the optimization landscape of (4) is too flat around some saddle points and hence not amenable to fast computation. It would be interesting to investigate what families of loss functions allow us to rule out bad local minima and eliminate the need of careful initialization, which might be better suited for tensor recovery problems.

Finally, in statistical inference and decision making, one might not be simply satisfied with obtaining a reliable estimate for each missing entry, but would also like to report a short confidence interval which is likely to contain the true entry. This boils down to the fundamental task of uncertainty quantification for tensor completion, which we leave to future investigation.

## Acknowledgments

The authors thank Lanqing Yu for many helpful discussions and Yuling Yan for proofreading the paper. This paper was first submitted in November 2019 and was accepted in part to NeurIPS 2019 (Cai et al. 2019).

## Endnotes

[1] We focus on symmetric order-3 tensors primarily for simplicity of presentation. Many of our findings naturally extend to the more general case with asymmetric tensors of possibly higher order. Detailed discussions are deferred to Section EC.7 in the e-companion due to the space limits.

[2] Here, a tensor $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$ is said to be *symmetric* if $T_{j,k,l} = T_{k,j,l} = T_{k,l,j} = T_{l,k,j} = T_{j,l,k} = T_{l,j,k}$ for all $1 \le j,k,l \le d$.

## References

Abbe E, Fan J, Wang K, Zhong Y (2017) Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* 48(3):1452–1474.

Anandkumar A, Ge R, Janzamin M (2014b) Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. Preprint, submitted February 21, https://arxiv.org/abs/1402.5180.

Anandkumar A, Ge R, Hsu D, Kakade SM, Telgarsky M (2014a) Tensor decompositions for learning latent variable models. *J. Machine Learn. Res.* 15(1):2773–2832.

Ashraphijuo M, Wang X (2017) Fundamental conditions for low-CP-rank tensor completion. *J. Machine Learn. Res.* 18(1):2116–2145.

Barak B, Moitra A (2016) Noisy tensor completion via the sum-of-squares hierarchy. *Proc. 29th Annual Conf. Learning Theory, New York, June 23–26*, 417–445.

Bubeck S (2015) Convex optimization: Algorithms and complexity. *Foundations Trends Machine Learning* 8(3–4):231–357.

Cai C, Li G, Poor HV, Chen Y (2019) Nonconvex low-rank tensor completion from noisy data. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Adv. Neural Inform. Processing Systems 32* (Curran Associates, Red Hook, NY), 1863–1874.

Cai C, Li G, Chi Y, Poor HV, Chen Y (2021) Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *Ann. Statist.* 49(2):944–967.

Candès E, Recht B (2009) Exact matrix completion via convex optimization. *Foundations Comput. Math.* 9(6):717–772.

Candès EJ, Li X, Soltanolkotabi M (2015) Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Trans. Inform. Theory* 61(4):1985–2007.

Chen Y, Candès EJ (2017) Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Comm. Pure Appl. Math.* 70(5):822–883.

Chen Y, Chi Y (2014) Robust spectral compressed sensing via structured matrix completion. *IEEE Trans. Inform. Theory* 60(10): 6576–6601.

Chen Y, Chi Y (2018) Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine* 35(4):14–31.

Chen Y, Wainwright MJ (2015) Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. Preprint, submitted September 10, https://arxiv.org/abs/1509.03025.

Chen J, Liu D, Li X (2019a) Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. Preprint, submitted January 18, https://arxiv.org/abs/1901.06116v1.

Chen H, Raskutti G, Yuan M (2019b) Non-convex projected gradient descent for generalized low-rank tensor regression. *J. Machine Learn. Res.* 20(1):172–208.

Chen Y, Chi Y, Fan J, Ma C (2019c) Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Math. Programming* 176(1–2):5–37.

Chen Y, Fan J, Ma C, Wang K (2019d) Spectral method and regularized MLE are both optimal for top-*K* ranking. *Ann. Statist.* 47 (4):2204–2235.

Chen Y, Fan J, Ma C, Yan Y (2019e) Inference and uncertainty quantification for noisy matrix completion. *Proc. Natl. Acad. Sci. USA* 116(46):22931–22937.

Chen Y, Chi Y, Fan J, Ma C, Yan Y (2021) Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM J. Optim.* 30(4):3098–3121.

Cheng JY, Zhang T, Alley MT, Uecker M, Lustig M, Pauly JM, Vasanawala SS (2017) Comprehensive multi-dimensional MRI for the simultaneous assessment of cardiopulmonary anatomy and physiology. *Sci. Rep.* 7(1):5330.

Chi Y, Lu Y, Chen Y (2019) Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Trans. Signal Processing* 67(20):5239–5269.

Davenport MA, Romberg J (2016) An overview of low-rank matrix recovery from incomplete observations. *IEEE J. Selected Topics Signal Processing* 10(4):608–622.

Ding L, Chen Y (2018) Leave-one-out approach for matrix completion: Primal and dual analysis. Preprint, submitted March 20, https://arxiv.org/abs/1803.07554.

Driggs D, Becker S, Boyd-Graber J (2019) Tensor robust principal component analysis: Better recovery with atomic norm regularization. Preprint, submitted January 30, https://arxiv.org/abs/1901.10991.

El Karoui N (2015) On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields* 170: 95–175.

Ely G, Aeron S, Hao N, Kilmer ME (2013) 5D and 4D pre-stack seismic data completion using tensor nuclear norm (TNN). *SEG Tech. Program Expanded Abstr.* (Society of Exploration Geophysicists, Tulsa, OK), 3639–3644.

Farias VF, Li AA (2019) Learning preferences with side information. *Management Sci.* 65(7):3131–3149.

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

Ge R, Ma T (2017) On the optimization landscape of tensor decompositions. von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, eds. *Adv. Neural Inform. Processing Systems* 31 (Curran Associates, Red Hook, NY), 3653–3663.

Ge R, Huang F, Jin C, Yuan Y (2015) Escaping from saddle points: Online stochastic gradient for tensor decomposition. *Proc. 28th Conf. Learning Theory, Paris*, July 3–6, 797–842.

Gilboa D, Buchanan S, Wright J (2018) Efficient dictionary learning with gradient descent. Preprint, submitted September 27, https://arxiv.org/abs/1809.10313.

Goldfarb D, Qin Z (2014) Robust low-rank tensor recovery: Models and algorithms. *SIAM J. Matrix Anal. Appl.* 35(1):225–253.

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

Hao B, Zhang A, Cheng G (2020) Sparse and low-rank tensor estimation via cubic sketchings. *IEEE Trans. Inform. Theory* 66(9): 5927–5964.

Hao B, Wang B, Wang P, Zhang J, Yang J, Sun WW (2019) Sparse tensor additive regression. Submitted March 31, https://arxiv.org/abs/1904.00479.

Hillar CJ, Lim LH (2013) Most tensor problems are np-hard. *J. ACM* 60(6):45.

Hopkins SB, Schramm T, Shi J, Steurer D (2016) Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. *Proc.48th Annual ACM Sympos. Theory Comput.* (ACM, New York), 178–191.

Huang B, Mu C, Goldfarb D, Wright J (2015) Provable models for robust low-rank tensor completion. *Pacific J. Optim.* 11(2): 339–364.

Jain P, Oh S (2014) Provable tensor factorization with missing data. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems* 27 (Curran Associates, Red Hook, NY), 1431–1439.

Ji TY, Huang TZ, Zhao XL, Ma TH, Liu G (2016) Tensor completion using total variation and low-rank matrix factorization. *Inform. Sci.* 326:243–257.

Kasai H, Mishra B (2016) Low-rank tensor completion: a riemannian manifold preconditioning approach. Balcan MF, Weinberger KQ, eds. *Proc. 33rd Internat. Conf. Machine Learning, New York*, June 20–22, 1012–1021.

Keshavan RH, Montanari A, Oh S (2010a) Matrix completion from a few entries. *IEEE Trans. Inform. Theory* 56(6):2980–2998.

Keshavan RH, Montanari A, Oh S (2010b) Matrix completion from noisy entries. *J. Machine Learn. Res.* 11(69):2057–2078.

Kilmer ME, Braman K, Hao N, Hoover RC (2013) Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM J. Matrix Anal. Appl.* 34(1):148–172.

Kim HJ, Ollila E, Koivunen V, Croux C (2013) Robust and sparse estimation of tensor decompositions. *Proc. 2013 IEEE Global Conf. Signal Inform. Processing* (IEEE, Piscataway, NJ), 965–968.

Kolda TG (2001) Orthogonal tensor decompositions. *SIAM J. Matrix Anal. Appl.* 23(1):243–255.

Kolda TG, Bader BW (2009) Tensor decompositions and applications. *SIAM Rev.* 51(3):455–500.

Kreimer N, Stanton A, Sacchi MD (2013) Tensor completion based on nuclear norm minimization for 5d seismic data reconstruction. *Geophysics* 78(6):273–284.

Krishnamurthy A, Singh A (2013) Low-rank matrix and tensor completion via adaptive sampling. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems* 26 (Curran Associates, Red Hook, NY), 836–844.

Li Q, Tang G (2017) Convex and nonconvex geometries of symmetric tensor factorization. *Proc. 51st Asilomar Conf. Signals Systems Comput.* (IEEE, Piscataway, NJ), 305–309.

Li X, Ye Y, Xu X (2017) Low-rank tensor completion with total variation for visual data inpainting. Singh S, Markovitch S, eds. *Proc. 31st AAAI Conf. Artificial Intelligence* (AAAI Press, San Francisco), 2210–2216.

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

Liu J, Musialski P, Wonka P, Ye J (2013) Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Anal. Machine Intelligence* 35(1):208–220.

Liu Y, Shang F, Cheng H, Cheng J, Tong H (2014) Factor matrix trace norm minimization for low-rank tensor completion. Zaki M, Obradovic Z, Tan PN, Banerjee A, Kamath C, Parthasarathy S, eds. *Proc. 2014 SIAM Internat. Conf. Data Mining* (SIAM, Philadelphia), 866–874.

Lounici K (2014) High-dimensional covariance matrix estimation with missing observations. *Bernoulli* 20(3):1029–1058.

Lu C, Feng J, Chen Y, Liu W, Lin Z, Yan S (2016) Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. *Proc. IEEE Conf. Comput. Vision Pattern Recognition* (IEEE, Piscataway, NJ), 5249–5257.

Ma C, Wang K, Chi Y, Chen Y (2017) Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *Foundations Comput. Math.* 20:451–463.

Mišić VV, Perakis G (2020) Data analytics in operations management: A review. *Manufacturing Service Oper. Management* 22(1): 158–169.

Montanari A, Sun N (2018) Spectral algorithms for tensor completion. *Comm. Pure Appl. Math.* 71(11):2381–2425.

Mu C, Huang B, Wright J, Goldfarb D (2014) Square deal: Lower bounds and improved relaxations for tensor recovery. Xing EP, Jebara T, eds. *Proc. 31st Internat. Conf. Machine Learning, Beijing*, June 22–24, 73–81.

Pananjady A, Wainwright MJ (2019) Instance-dependent $\ell_\infty$-bounds for policy evaluation. Preprint, submitted September 19, https://arxiv.org/abs/1909.08749.

Pawlowski C (2019) Machine learning for problems with missing and uncertain data with applications to personalized medicine. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Potechin A, Steurer D (2017) Exact tensor completion with sum-of-squares. Kale S, Shamir O, eds. *Proc. 2017 Conf. Learning Theory, Amsterdam*, July 7–10, 1619–1673.

Rauhut H, Schneider R, Stojanac Ž (2017) Low rank tensor recovery via iterative hard thresholding. *Linear Algebra Appl.* 523:220–262.

Richard E, Montanari A (2014) A statistical model for tensor PCA. Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Adv. Neural Inform. Processing, Systems* 27 (Curran Associates, Red Hook, NY), 2897–2905.

Romera-Paredes B, Pontil M (2013) A new convex relaxation for tensor completion. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems* 26 (Curran Associates, Red Hook, NY), 2967–2975.

Semerci O, Hao N, Kilmer ME, Miller EL (2014) Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. *IEEE Trans. Image Processing* 23(4): 1678–1693.

Shah D, Yu CL (2019) Iterative collaborative filtering for sparse noisy tensor estimation. Preprint, submitted August 3, https://arxiv.org/abs/1908.01241.

Sidiropoulos ND, De Lathauwer L, Fu X, Huang K, Papalexakis EE, Faloutsos C (2017) Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Processing* 65(13): 3551–3582.

Soroushmehr SR, Najarian K (2016) Transforming big data into computational models for personalized medicine and healthcare. *Dialogues Clinical Neuroscience* 18(3):339–343.

Steinlechner M (2016) Riemannian optimization for high-dimensional tensor completion. *SIAM J. Sci. Comput.* 38(5):S461–S484.

Sun R, Luo ZQ (2016) Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inform. Theory* 62(11):6535–6579.

Sun WW, Lu J, Liu H, Cheng G (2017) Provable sparse tensor decomposition. *J. Royal Statist. Soc. Ser. B. Statist. Methodology* 79(3): 899–916.

Tan YS, Vershynin R (2019) Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. Preprint, submitted October 28, https://arxiv.org/abs/1910.12837.

Tang G, Shah P (2015) Guaranteed tensor decomposition: A moment approach. Bach F, Blei D, eds. *Proc. 32nd Internat. Conf. Machine Learning, Lille, France*, July 7–9, 1491–1500.

Tomioka R, Hayashi K, Kashima H (2010) Estimation of low-rank tensors via convex optimization. Preprint, submitted October 5, https://arxiv.org/abs/1010.0789.

Wang W, Aggarwal V, Aeron S (2016) Tensor completion by alternating minimization under the tensor train (TT) model. Preprint, submitted September 19, https://arxiv.org/abs/1609.05587.

Xia D, Yuan M (2017) On polynomial time methods for exact low rank tensor completion. Preprint, submitted February 22, https://arxiv.org/abs/1702.06980.

Xia D, Yuan M, Zhang CH (2017) Statistically optimal and computationally efficient low rank tensor completion from noisy entries. Preprint, submitted November 14, https://arxiv.org/abs/1711.04934.

Xu Y, Yin W (2013) A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.* 6(3): 1758–1789.

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

Yao Q (2018) Scalable tensor completion with nonconvex regularization. Preprint, submitted July 23, https://arxiv.org/abs/1807.08725v1.

Yi X, Park D, Chen Y, Caramanis C (2016) Fast algorithms for robust PCA via gradient descent. Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, eds. *Adv. Neural Inform. Processing Systems* 29 (Curran Associates, Red Hook, NY), 4152–4160.

Ying J, Lu H, Wei Q, Cai JF, Guo D, Wu J, Chen Z, Qu X (2017) Hankel matrix nuclear norm regularized tensor completion for $n$-dimensional exponential signals. *IEEE Trans. Signal Processing* 65(14):3702–3717.

Yuan M, Zhang CH (2016) On tensor completion via nuclear norm minimization. *Found. Comput. Math.* 16(4):1031–1068.

Yuan M, Zhang CH (2017) Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Trans. Inform. Theory* 63(10):6753–6766.

Zhang A (2019) Cross: Efficient low-rank tensor completion. *Ann. Statist.* 47(2):936–964.

Zhang Z, Aeron S (2017) Exact tensor completion using t-SVD. *IEEE Trans. Signal Processing* 65(6):1511–1526.

Zhang A, Xia D (2018) Tensor SVD: Statistical and computational limits. *IEEE Trans. Inform. Theory* 64(11):7311–7338.

Zhong Y, Boumal N (2018) Near-optimal bound for phase synchronization. *SIAM J. Optim.* 28(2):989–1016.

Zoubir AM, Koivunen V, Ollila E, Muma M (2018) *Robust Statistics for Signal Processing* (Cambridge University Press, Cambridge, UK).

**Changxiao Cai** is a PhD student in the Department of Electrical Engineering at Princeton University. He received the BE in electronic engineering from Tsinghua University in 2016. His research interests include machine learning, high-dimensional statistics, optimization, and information theory.

**Gen Li** is a PhD student in the Department of Electronic Engineering at Tsinghua University. He received the BE degree in electronic engineering from Tsinghua University in 2016. His recent research interests include machine learning and nonconvex optimization, high-dimensional statistics, and reinforcement learning.

**H. Vincent Poor** is the Michael Henry Strater University Professor of Electrical Engineering at Princeton University. His research interests are in the areas of information theory, machine learning, and network science and their applications in wireless networks, energy systems, and related fields. Recent recognition of his

work includes the 2017 IEEE Alexander Graham Bell Medal and a DEng honoris causa from the University of Waterloo awarded in 2019.

**Y. Chen** is an assistant professor in the Department of Electrical Engineering at Princeton University. His research interests include high-dimensional statistics, convex and nonconvex optimization, reinforcement learning, and information theory. His most recent awards include the 2020 Army Research Office Young Investigator Award, the 2020 Princeton Graduate Mentoring Award, and the 2020 International Congress of Chinese Mathematicians Best Paper Award (gold medal).