AN OPTIMAL STATISTICAL AND COMPUTATIONAL FRAMEWORK FOR GENERALIZED TENSOR ESTIMATION

BY RUNGANG HAN1,*, REBECCA WILLETT2, AND ANRU R. ZHANG1,†

1Department of Statistics, University of Wisconsin-Madison, *rhan32@stat.wisc.edu; 1anruzhang@stat.wisc.edu
2Departments of Statistics and Computer Science, University of Chicago, willett@uchicago.edu

This paper describes a flexible framework for generalized low-rank tensor estimation problems that includes many important instances arising from applications in computational imaging, genomics, and network analysis. The proposed estimator consists of finding a low-rank tensor fit to the data under generalized parametric models. To overcome the difficulty of non-convexity in these problems, we introduce a unified approach of projected gradient descent that adapts to the underlying low-rank structure. Under mild conditions on the loss function, we establish both an upper bound on statistical error and the linear rate of computational convergence through a general deterministic analysis. Then we further consider a suite of generalized tensor estimation problems, including sub-Gaussian tensor PCA, tensor regression, and Poisson and binomial tensor PCA. We prove that the proposed algorithm achieves the minimax optimal rate of convergence in estimation error. Finally, we demonstrate the superiority of the proposed framework via extensive experiments on both simulated and real data.

1. Introduction. In recent years, the analysis of tensors or high-order arrays has emerged as an active topic in statistics, applied mathematics, machine learning, and data science. Datasets in the form of tensors arise from various scientific applications (Kroonenberg, 2008), such as collaborative filtering (Bi, Qu and Shen, 2018; Shah and Yu, 2019), neuroimaging analysis (Zhou, Li and Zhu, 2013; Li et al., 2018), hyperspectral imaging (Li and Li, 2010), longitudinal data analysis (Hoff, 2015), and more. In many of these problems, although the tensor of interest is high-dimensional in the sense that the ambient dimension of the dataset is substantially greater than the sample size, there is often hidden low-dimensional structures in the tensor that can be exploited to facilitate the data analysis. In particular, the low-rank condition renders convenient decomposable structure and has been proposed and widely used in the analysis of tensor data (Kroonenberg, 2008; Kolda and Bader, 2009). However, leveraging these hidden low-rank structures in estimation and inference can pose great statistical and computational challenges in real practice.

1.1. Generalized Tensor Estimation. In this paper, we consider a statistical and optimization framework for generalized tensor estimation. Suppose we observe a random sample D drawn from some distribution parametrized by an unknown low-rank tensor parameter \( X^* \in \mathbb{R}^{p_1 \times p_2 \times p_3} \). A straightforward idea to estimate \( X^* \) is via optimization:

\[
\hat{X} = \arg \min_{X \text{ is low-rank}} L(X; D).
\]

Here, \( L(X; D) \) can be taken as the negative log-likelihood function (then \( \hat{X} \) becomes the maximum likelihood estimator (MLE)) or any more general loss function. We can even

MSC2020 subject classifications: Primary 62H12, 62H25; secondary 62C20.
Keywords and phrases: generalize tensor estimation, gradient descent, image denoising, low-rank tensor, minimax optimality, non-convex optimization.
broaden the scope of this framework to a deterministic setting: suppose we observe \( D \) that is “associated” with an unknown tensor parameter \( \mathcal{X}^* \in \mathbb{R}^{p_1 \times p_2 \times p_3} \); to estimate \( \mathcal{X}^* \), we try to minimize the loss function \( L(\mathcal{X}; D) \) that is specified by the problem scenario. This general framework includes many important instances arising in real applications. For example:

- **Computational imaging.** Photon-limited imaging appears in signal processing (Salmon et al., 2014), material science (Yankovich et al., 2016), astronomy (Timmerman and Nowak, 1999; Willett and Nowak, 2007), and often involves arrays with non-negative photon counts contaminated by substantial noise. Data from photon-limited imaging are often in the form of tensors (e.g., stacks of spectral images in which each image corresponds to a different wavelength of light). How to denoise these images is often crucial for the subsequent analysis. To this end, Poisson tensor PCA serves as a prototypical model for tensor photon-limited imaging analysis; see Sections 4.3 and 7.2 for more details.

- **Analysis of multilayer network data.** In network analysis, one often observes multiple snapshots of static or dynamic networks (Sewell and Chen, 2015; Lei, Chen and Lynch, 2019; Arroyo et al., 2019; Pensky et al., 2019). How to perform an integrative analysis for the network structure using multilayer network data has become an important problem in practice. By stacking adjacency matrices from multiple snapshots to an adjacency tensor, the hidden community structure of network can be transformed to the low-rankness of adjacency tensor, and the generalized tensor learning framework can provide a new perspective on the analysis of multilayer network data.

- **Biological sequencing data analysis.** Tensor data also commonly appear in biological sequencing data analysis (Faust et al., 2012; Flores et al., 2014; Wang, Fischer and Song, 2017). The identification of significant triclusters or modules, i.e., coexpressions of different genes or coexistence of different microbes, often has significant biological meanings (Henriques and Madeira, 2019). From a statistical perspective, these modules often correspond to low-rank tensor structure, so the generalized tensor learning framework could be naturally applied.

- **Online-click through Prediction.** Online click-through data analysis in e-commerce has become an increasingly important tool in building the online recommendation system (McMahan et al., 2013; Sun and Li, 2016; Shan et al., 2016). There are three major entities: users, items, and time, and the data can be organized as a tensor, where each entry represents the click times of one user on a specific category of items in a time period (e.g., noon or evening). Then generalized tensor estimation could be applied to study the implicit features of users and items for better prediction of user behaviors.

Additional applications include neuroimaging analysis (Zhou, Li and Zhu, 2013), collaborative filtering (Yu et al., 2018), mortality rate analysis (Wilmoth and Shkolnikov, 2006), and more. We also elucidate specific model setups and real data examples in detail later in Sections 4 and 7.2, respectively.

The central tasks of generalized tensor estimation problems include two elements. From a statistical perspective, it is important to investigate how well one can estimate the target tensor parameter \( \mathcal{X}^* \) and the optimal rates of estimation error. From an optimization perspective, it is crucial to develop a computationally efficient procedure for estimating \( \mathcal{X}^* \) with provable theoretical guarantees. To estimate the low-rank tensor parameter \( \mathcal{X}^* \), a straightforward idea is to perform the rank constrained minimization on the loss function \( L(\mathcal{X}; D) \) in (1.1). Since the low-rank constraint is highly non-convex, the direct implementation of (1.1) is computationally infeasible in practice. If \( \mathcal{X}^* \) is a sparse vector or low-rank matrix, common substitutions often involve convex regularization methods, such as M-estimators with an \( \ell_1 \) penalty or matrix nuclear norm penalty for estimating sparse or low-rank structure (Tibshirani, 1996; Fazel, 2002). These methods enjoy great empirical and theoretical success.
for vector/matrix estimators, but it is unclear whether they can achieve good performance on generalized tensor estimation problems. First, different from the matrix nuclear norm, tensor nuclear norm is generally NP-hard to even approximate (Friedland and Lim, 2018), so that the tensor nuclear norm regularization approach can be computationally intractable. Second, other computationally feasible convex regularization methods, such as the overlapped nuclear norm minimization (Tomioka et al., 2011; Tomioka and Suzuki, 2013), may be statistical sub-optimal based on the theory of simultaneously structured model estimation (Oymak et al., 2015).

In contrast, we focus on a unified non-convex approach for generalized tensor estimation problems in this paper. Our central idea is to decompose the low-rank tensor into $\mathcal{X} = [S; U_1, U_2, U_3]$ (see Section 2.1 for explanations of tensor algebra) and reformulate the original problem to

$$
(\hat{S}, \hat{U}_1, \hat{U}_2, \hat{U}_3) = \arg \min_{S, U_1, U_2, U_3} \left\{ L([S; U_1, U_2, U_3]; D) + \frac{a}{2} \sum_{k=1}^{3} \| U_k^\top U_k - b^2 I_{r_k} \|_F^2 \right\},
$$

which can be efficiently solved by (projected) gradient descent on all components. The resulting $\mathcal{X} = [\hat{S}; \hat{U}_1, \hat{U}_2, \hat{U}_3]$ naturally admits a low-rank structure. The auxiliary regularizers $\| U_k^\top U_k - b^2 I_{r_k} \|_F^2$ in (1.2) can keep $\hat{U}_k$ from being singular. It is actually easy to check that (1.1) and (1.2) are exactly equivalent.

We provide strong theoretical guarantees for the proposed procedure on generalized tensor estimation problems. In particular, we establish the linear rate of local convergence for gradient descent methods under a general deterministic setting with the Restricted Correlated Gradient condition (see Section 3.1 for details). An informal statement of the result is given below,

$$
\| \mathcal{X}(t) - \mathcal{X}^* \|_F^2 \lesssim \xi^2 + (1 - c)^t \| \mathcal{X}(0) - \mathcal{X}^* \|_F^2
$$

for all $t \geq 1$ with high probability. Here, we use $\xi^2$ to characterize the statistical noise and its definition and interpretation will be given in section 3.2. Then for specific statistical models, including sub-Gaussian tensor PCA, tensor regression, Poisson tensor PCA, and binomial tensor PCA, based on the general result (1.3), we prove that the proposed algorithm achieves the minimax optimal rate of convergence in estimation error. Specifically for the low-rank tensor regression problem, Table 1 illustrates the advantage of our method through a comparison with existing ones.

Finally, we apply the proposed framework to synthetic and real data examples, including photon-limited 4D-STEM (scanning transmission electron microscopy) imaging data and click-through e-commerce data. The comparison of performance with existing methods illustrates the merit of our proposed procedure.

1.2. Related Literature. This work is related to a broad range of literature on tensor analysis. For example, tensor decomposition/SVD/PCA focuses on the extraction of low-rank structures from noisy tensor observations (Richard and Montanari, 2014; Anandkumar et al., 2014; Hopkins, Shi and Steurer, 2015; Montanari, Reichman and Zeitouni, 2017; Lesieur et al., 2017; Johndrow, Bhattacharya and Dunson, 2017; Zhang and Xia, 2018; Chen, 2019). Correspondingly, a number of methods have been proposed and analyzed under either deterministic or random Gaussian noise, such as the maximum likelihood estimation (Richard and

---

\footnote{The analysis in Rauhut, Schneider and Stojanac (2017) relies on an assumption that the projection on low-rank tensor manifold can be approximately done by High-Order SVD. It is, however, unclear whether this assumption holds in general.}
Montanari, 2014), (truncated) power iterations (Anandkumar et al., 2014; Sun et al., 2017), higher-order SVD (De Lathauwer, De Moor and Vandewalle, 2000a; Zhang and Xia, 2018), higher-order orthogonal iteration (HOOI) (De Lathauwer, De Moor and Vandewalle, 2000b; Zhang and Xia, 2018), STAT-SVD (Zhang and Han, 2018).

Since non-Gaussian-valued tensor data also commonly appear in practice, Signoretto et al. (2011); Chi and Kolda (2012); Hong, Kolda and Duersch (2018) considered the generalized tensor decomposition and introduced computational efficient algorithms. However, the theoretical guarantees for these procedures and the statistical limits of the generalized tensor decomposition still remain open.

Our proposed framework includes the topic of tensor recovery and tensor regression. Various methods, such as the convex regularization (Tomioka and Suzuki, 2013; Raskutti et al., 2019), alternating minimization (Zhou, Li and Zhu, 2013), hard thresholding iteration (Chen, Raskutti and Yuan, 2019; Rauhut, Schneider and Stojanac, 2017, 2015), importance-sketching (Zhang et al., 2019) were introduced and studied. A more detailed comparison of these methods is summarized in Table 1.

In addition, high-order interaction pursuits (Hao, Zhang and Cheng, 2019), tensor completion (Liu et al., 2013; Yuan and Zhang, 2014; Montanari and Sun, 2018; Xia and Yuan, 2017; Xia, Yuan and Zhang, 2017; Zhang, 2019; Cai et al., 2019), and tensor block models (Chi et al., 2018; Lei, Chen and Lynch, 2019; Wang and Zeng, 2019) are important topics in tensor analysis that have attracted enormous attention recently. Departing from the existing results, this paper, to the best of our knowledge, is the first to give a unified treatment for a broad range of tensor estimation problems with both statistical optimality and computational efficiency.

This work is also related to a substantial body of literature on low-rank matrix recovery, where the goal is to estimate a low-rank matrix based on a limited number of observations. Specific examples of this topic include matrix completion (Candès and Recht, 2009; Candès and Plan, 2010), phase retrieval (Candès, Li and Soltanolkotabi, 2015; Cai, Li and Ma, 2016), blind deconvolution (Ahmed, Recht and Romberg, 2013), low-rank matrix trace regression (Keshavan, Montanari and Oh, 2010; Koltchinskii, Lounici and Tsybakov, 2011; Chen and Chi, 2018; Fan, Gong and Zhu, 2019), and many others. A common approach for low-rank matrix recovery is via explicit low-rank factorization: one can decompose the

**Table 1**

| Algorithm                       | Sample complexity* | Estimation Error Upper Bound | Recovery (noiseless) |
|--------------------------------|--------------------|------------------------------|----------------------|
| Our Method                      | $p^{3/2}r$         | $\sigma^2 pr/n$              | Exact                |
| Tucker-Reg. (Zhou, Li and Zhu, 2013) | N.A.               | N.A.                         | Exact                |
| Nonconvex-PGD. (Chen, Raskutti and Yuan, 2019) | $p^2r$             | $\sigma^2 p^2r/n$            | Exact                |
| Nuclear Norm Min. (Raskutti et al., 2019) | N.A.               | $\sigma^2 pr^2/n$            | Exact                |
| Schatten-1 Norm Min. (Tomioka and Suzuki, 2013) | $p^2r$             | $\sigma^2 p^2r/n$            | Exact                |
| ISLET (Zhang et al., 2019)      | $p^{3/2}r$         | $\sigma^2 pr/n$              | Inexact              |
| Iterative Hard Thresholding† (Rauhut, Schneider and Stojanac, 2017) | $pr$               | $\sigma^2$                   | Exact                |

Comparison of different tensor regression methods when the rank is known. For simplicity, we assume $r_1 = r_2 = r_3 = r$, $p_1 = p_2 = p_3 = p$ and $\sigma^2 \ll \|X^*\|_F^2$. Here, the sample complexity* is the minimal sample size required to achieve the corresponding estimation error.
target $p_1$-by-$p_2$ rank-$r$ matrix $X$ into $X = UV^\top$, where $U \in \mathbb{R}^{p_1 \times r}$, $V \in \mathbb{R}^{p_2 \times r}$, then minimize the loss function $L(UV^\top)$ with respect to both $U$ and $V$ (Wen, Yin and Zhang, 2012). Previously, Zhao, Wang and Liu (2015) considered the noiseless setting of trace regression and proved that under good initialization, the first order alternating optimization on $U$ and $V$ achieves exact recovery. Tu et al. (2016); Park et al. (2018) established the local convergence of gradient descent for strongly convex and smooth loss function $L$. The readers are referred to a recent survey paper (Chi, Lu and Chen, 2019) on the applications and optimization landmarks of the non-convex factorized optimization. Despite significant developments in low-rank matrix recovery and non-convex optimization, they cannot be directly generalized to tensor estimation problems for many reasons. First, many basic matrix concepts or methods cannot be directly generalized to high-order ones (Hillar and Lim, 2013). Naive generalization of matrix concepts (e.g., operator norm, singular values, eigenvalues) are possible but often computationally NP-hard. Second, tensors have more complicated algebraic structure than matrices. As what we will illustrate later, one has to simultaneously handle all arm matrices (i.e., $U_1$, $U_2$, and $U_3$) and the core tensor (i.e., $S$) with distinct dimensions in the theoretical error contraction analysis. To this end, we develop new technical tools on tensor algebra and perturbation results (e.g., E.2, Lemmas E.3 in the Appendix). More technical issues of generalized tensor estimation will be addressed in Section 3.3.

The projected gradient schemes, which apply gradient descent on the parameter tensor $X$ followed by the low-rank tensor retraction/projection operators, form another important class of methods in the literature (Rauhut, Schneider and Stojanac, 2015, 2017; Chen, Raskutti and Yuan, 2019):

$$X^{(t+1)} = P \left( X^{(t)} - \eta \nabla L(X^{(t)}) \right).$$

Different from the low-rank projection for matrices, the exact low-rank tensor projection (i.e., the best rank-$(r_1, r_2, r_3)$ approximation: $P(X) = \arg\min_T \|T - X\|_F$) is NP-hard in general (Hillar and Lim, 2013) and less practical. Several inexact but efficient projection methods were developed and studied to overcome this issue. In particular, Rauhut, Schneider and Stojanac (2015) proposed a polynomial-time computable projected gradient scheme that converges linearly to the true tensor parameter for the noiseless tensor completion problem, given the initialization is sufficiently close to the solution. On the other hand, it is not clear if such schemes with inexact projection operators can achieve optimal statistical rate in the noisy setting (Chen, Raskutti and Yuan, 2019). In contrast, the proposed method in this paper is both computationally efficient and statistically optimal in a variety of settings with provable guarantees.

1.3. Organization of the Paper. The rest of the article is organized as follows. After a brief introduction of the notation and preliminaries in Section 2.1, we introduce the general problem formulation in Section 2.2. A deterministic error and local convergence analysis of the projected gradient descent algorithm for order-3 tensor estimation is discussed in Section 3. Then we apply the results on a variety of generalized tensor estimation problems in Section 4, including sub-Gaussian tensor PCA, tensor regression, Poisson tensor PCA, and binomial tensor PCA. We develop the upper and minimax matching lower bounds in each of these scenarios. In Section 5, we propose a data-driven rank selection method with theoretical guarantee. The extension to general order-$d$ tensor estimation is discussed in Section 6. Simulation and real data analysis are presented in Section 7. All proofs of technical results and more implementation details of algorithms are collected in the supplementary materials.

2. Generalized Tensor Estimation Model.
2.1. Notation and Preliminaries. The following notation and preliminaries are used throughout this paper. The lowercase letters, e.g., \(x, y, u, v\), are used to denote scalars or vectors. For any \(a, b \in \mathbb{R}\), let \(a \wedge b\) and \(a \vee b\) be the minimum and maximum of \(a\) and \(b\), respectively. We use \(C, C_0, C_1, \ldots\) and \(c, c_0, c_1, \ldots\) to represent generic large and small positive constants respectively. The actual values of these generic symbols may differ from line to line.

We use bold uppercase letters \(\mathbf{A}, \mathbf{B}\) to denote matrices. Let \(\mathbb{O}_{p,r}\) be the collection of all \(p\)-by-\(r\) matrices with orthonormal columns: \(\mathbb{O}_{p,r} = \{\mathbf{U} \in \mathbb{R}^{p \times r} : \mathbf{U}^\top \mathbf{U} = \mathbf{I}_r\}\), where \(\mathbf{I}_r\) is the \(r\)-by-\(r\) identity matrix. For any matrix \(\mathbf{A} \in \mathbb{R}^{p_1 \times p_2}\), let \(\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_{\min(p_1, p_2)}(\mathbf{A}) \geq 0\) be its singular values in descending order. We also define \(\text{SVD}_r(\mathbf{A}) \in \mathbb{O}_{p,r}\) to be the matrix comprised of the top \(r\) left singular vectors of \(\mathbf{A}\). For any matrix \(\mathbf{A}\), let \(\mathbf{A}_{ij}, \mathbf{A}_{ij},\) and \(\mathbf{A}_{ij}\) be the entry on the \(i\)th row and \(j\)th column, the \(i\)th row, and the \(j\)th column of \(\mathbf{A}\), respectively.

The inner product of two matrices with the same dimension is defined as \(\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^\top \mathbf{B})\), where \(\text{tr}(\cdot)\) is the trace operator. We use \(\|\mathbf{A}\|_F = \sigma_1(\mathbf{A})\) to denote the spectral norm of \(\mathbf{A}\), use \(\|\mathbf{A}\|_2 = \sqrt{\sum_{i,j} A_{ij}^2}\) to denote the Frobenius norm of \(\mathbf{A}\), and use \(\|\mathbf{A}\|_* = \sum_{k=1}^{p_1 \wedge p_2} \sigma_k\) to denote the nuclear norm of \(\mathbf{A}\). The \(l_{2,\infty}\) norm of \(\mathbf{A}\) is defined as the largest row-wise \(l_2\) norm of \(\mathbf{A}\): \(\|\mathbf{A}\|_{l_{2,\infty}} = \max_i \|\mathbf{A}_{i,:}\|_2\). For any matrix \(\mathbf{A} = [a_1, \ldots, a_J] \in \mathbb{R}^{p \times J}\) and \(\mathbf{B} \in \mathbb{R}^{K \times L}\), the Kronenecker product is defined as the \((IK)\)-by-(\(JL\)) matrix \(\mathbf{A} \otimes \mathbf{B} = [a_1 \otimes \mathbf{B} \ldots a_J \otimes \mathbf{B}]\).

In addition, we use calligraphic letters, e.g., \(\mathcal{S}, \mathcal{X}, \mathcal{Y}\), to denote higher-order tensors. To simplify the presentation, we mainly focuses on order-3 tensors in this paper while all results for higher-order tensors can be carried out similarly. For tensor \(\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}\) and matrix \(\mathbf{U}_1 \in \mathbb{R}^{p_1 \times r_1}\), the mode-1 tensor-matrix product is defined as:

\[
\mathcal{S} \times_1 \mathbf{U}_1 = \mathbb{R}^{p_1 \times r_2 \times r_3}, \quad (\mathcal{S} \times_1 \mathbf{U}_1)_{ij_1 j_2 j_3} = \sum_{j=1}^{r_1} S_{j j_2 j_3} (\mathbf{U}_1)_{i j_1}.
\]

For any \(\mathbf{U}_2 \in \mathbb{R}^{p_2 \times r_2}, \mathbf{U}_3 \in \mathbb{R}^{p_3 \times r_3}\), the tensor-matrix products \(\mathcal{S} \times_2 \mathbf{U}_2\) and \(\mathcal{S} \times_3 \mathbf{U}_3\) are defined in a similarly way. Importantly, multiplication along different directions is commutative invariant: \((\mathcal{S} \times_{k_1} U_{k_1}) \times_{k_2} U_{k_2} = (\mathcal{S} \times_{k_2} U_{k_2}) \times_{k_1} U_{k_1}\) for any \(k_1 \neq k_2\). We simply denote

\[
\mathcal{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 = [\mathcal{S}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3],
\]

as this formula commonly appears in the analysis. We also introduce the matricization operator that transforms tensors to matrices: for \(\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}\), define

\[
\mathcal{M}_1(\mathcal{X}) = \mathbb{R}^{p_1 \times p_2 p_3}, \quad \text{where} \quad [\mathcal{M}_1(\mathcal{X})]_{i_1 i_2 p_3 (i_3 - 1)} = \mathcal{X}_{i_1 i_2 i_3},
\]

\[
\mathcal{M}_2(\mathcal{X}) = \mathbb{R}^{p_2 \times p_1 p_3}, \quad \text{where} \quad [\mathcal{M}_2(\mathcal{X})]_{i_2 i_3 p_1 (i_1 - 1)} = \mathcal{X}_{i_1 i_2 i_3},
\]

\[
\mathcal{M}_3(\mathcal{X}) = \mathbb{R}^{p_3 \times p_1 p_2}, \quad \text{where} \quad [\mathcal{M}_3(\mathcal{X})]_{i_3 i_1 p_2 (i_2 - 1)} = \mathcal{X}_{i_1 i_2 i_3},
\]

and \(\mathcal{M}_k^{-1} : \mathbb{R}^{p_k \times (p_{k+1} p_{k+2})} \rightarrow \mathbb{R}^{p_1 \times p_2 \times p_3}\) as the inverse operator of \(\mathcal{M}_k\) where \(k + 1\) and \(k + 2\) are computed modulo 3. Essentially, \(\mathcal{M}_k\) “flattens” all but the \(k\)th directions of any tensor. The following identity that relates the matrix-tensor product and matricization plays an important role in our analysis:

\[
\mathcal{M}_k(\mathcal{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3) = \mathbf{U}_k \mathcal{M}_k(\mathcal{S}) (\mathbf{U}_{k+2} \otimes \mathbf{U}_{k+1})^\top, \quad k = 1, 2, 3.
\]

Here again, \(k + 1\) and \(k + 2\) are computed modulo 3. The inner product of two tensors with the same dimension is defined as \(\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i,j,k} \mathcal{X}_{ijk} \mathcal{Y}_{ijk}\). The Frobenius norm of a tensor \(\mathcal{X}\) is defined as \(\|\mathcal{X}\|_F = \sqrt{\sum_{i,j,k} \mathcal{X}_{ijk}^2}\). For any smooth tensor-variate function \(f : \mathbb{R}^{p_1 \times p_2 \times p_3} \rightarrow \mathbb{R}\)
\( R \), let \( \nabla f : \mathbb{R}^{p_1 \times p_2 \times p_3} \to \mathbb{R}^{p_1 \times p_2 \times p_3} \) be the gradient function such that \( (\nabla f(X))_{ijk} = \frac{\partial f}{\partial X_{ijk}} \).

We simply write this as \( \nabla f \) when there is no confusion. Finally, the readers are also referred to Kolda and Bader (2009) for a comprehensive discussions on tensor algebra. The focus of this paper is on the following low-Tucker-rank tensors:

**Definition 2.1 (Low Tucker Rank)** We say \( X^* \in \mathbb{R}^{p_1 \times p_2 \times p_3} \) is Tucker rank-\((r_1, r_2, r_3)\) if and only if \( X^* \) can be decomposed as

\[
X^* = S^* \times_1 U_1^* \times_2 U_2^* \times_3 U_3^* =: [S^*; U_1^*, U_2^*, U_3^*]
\]

for some \( S^* \in \mathbb{R}^{r_1 \times r_2 \times r_3} \) and \( U_k^* \in \mathbb{R}^{p_k \times r_k}, k = 1, 2, 3. \)

In addition, \( X^* \) is Tucker rank-\((r_1, r_2, r_3)\) if and only if \( \text{rank}(M_k(X^*)) \leq r_k \) for \( k = 1, 2, 3 \).

For convenience of presentation, we denote \( \hat{p} = \max\{p_1, p_2, p_3\}, \hat{r} = \max\{r_1, r_2, r_3\}, p = \min\{p_1, p_2, p_3\}, \ell = \min\{r_1, r_2, r_3\}, p_k = p_1 p_2 p_3 / p_k \) and \( r_k = r_1 r_2 r_3 / r_k. \)

### 2.2. Generalized Tensor Estimation

Suppose we observe a dataset \( D \) associated with an unknown parameter \( X^* \). Here, \( X^* \) is a \( p_1 \)-by-\( p_2 \)-by-\( p_3 \) rank-\((r_1, r_2, r_3)\) tensor and \( r_k \ll p_k \). For example, \( D \) can be a random sample drawn from some distribution parametrized by \( X^* \) or a deterministic perturbation of \( X^* \). The central goal of this paper is to have an efficient and accurate estimation of \( X^* \).

Let \( L(X; D) \) be an empirical loss function known a priori, such as the negative log-likelihood function from the generating distribution or more general objective function. Then the following rank constrained optimization provides a straightforward way to estimate \( X^* \):

\[
\min_{X} L(X; D) \quad \text{subject to} \quad \text{rank}(M_k(X)) \leq r_k, \quad k = 1, 2, 3.
\]

As mentioned earlier, this framework includes many instances arising from applications in various fields. Due to the connection between low Tucker rank and the decomposition discussed in Section (2.1), it is natural to consider the following minimization problem

\[
\hat{X} = \hat{S} \times_1 \hat{U}_1 \times_2 \hat{U}_2 \times_3 \hat{U}_3,
\]

\[
(\hat{S}, \hat{U}_1, \hat{U}_2, \hat{U}_3) = \arg\min_{S, U_1, U_2, U_3} L(S \times_1 U_1 \times_2 U_2 \times_3 U_3; D),
\]

and consider a gradient-based optimization algorithm to estimate \( X^* \). Let \( \nabla L : \mathbb{R}^{p_1 \times p_2 \times p_3} \to \mathbb{R}^{p_1 \times p_2 \times p_3} \) be the gradient of loss function. The following Lemma gives the partial gradients of \( L \) on \( U_k \) and \( S \). The proof is provided in the supplementary material (Appendix E.1).

**Lemma 2.1 (Partial Gradients of Loss)**

\[
\nabla U_1 L ([S; U_1, U_2, U_3]) = M_1(\nabla L)(U_3 \otimes U_2)M_1(S)^\top,
\]

\[
\nabla U_2 L ([S; U_1, U_2, U_3]) = M_2(\nabla L)(U_1 \otimes U_3)M_2(S)^\top,
\]

\[
\nabla U_3 L ([S; U_1, U_2, U_3]) = M_3(\nabla L)(U_2 \otimes U_1)M_3(S)^\top,
\]

\[
\nabla S L ([S; U_1, U_2, U_3]) = \nabla L \times_1 U_1^\top \times_2 U_2^\top \times_3 U_3^\top.
\]

Here, \( \nabla L \) is short for \( \nabla L ([S; U_1, U_2, U_3]) \).

As mentioned earlier, we consider optimizing the following objective function:

\[
F(S, U_1, U_2, U_3) = L ([S; U_1, U_2, U_3]; D) + \frac{a}{2} \sum_{k=1}^{3} \left\| U_k^\top U_k - b^2 I_{r_k} \right\|_F^2.
\]
where \(a \geq 0, b > 0\) are tuning parameters to be discussed later. By adding regularizers \(\|U_k^\top U_k - b^2 I_{r_k}\|^2\), we can prevent \(U_k\) from being singular throughout gradient descent, while do not alter the minimizer. This can be summarized as the following proposition, whose proof is provided in Appendix E.2.

**Proposition 2.1** Suppose \((\hat{S}, \hat{U}_1, \hat{U}_2) = \arg\min F(S, U_1, U_2, U_3)\) for \(F\) defined in (2.4). Then
\[
\hat{x} = [\hat{S}, \hat{U}_1, \hat{U}_2] = \arg\min_{X: \text{rank}(X) \leq (r_1, r_2, r_3)} L(X; D).
\]

Similar regularizers have been widely used on non-convex low-rank matrix optimization (Tu et al., 2016; Park et al., 2018) and more technical interpretations are provided in Section 3.3.

### 3. Projected Gradient Descent

In this section, we study the local convergence of the projected gradient descent under a general deterministic framework.

#### 3.1. Restricted Correlated Gradient Condition

We first introduce the regularity condition on the loss function \(L\) and set \(\mathcal{C}\).

**Definition 3.1 (Restricted Correlated Gradient (RCG))** Let \(f\) be a real-valued function. We say \(f\) satisfies RCG \((\alpha, \beta, \mathcal{C})\) condition for \(\alpha, \beta > 0\) and the set \(\mathcal{C}\) if
\[
(3.1) \quad \langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle \geq \alpha \|x - x^*\|^2 + \beta \|\nabla f(x) - \nabla f(x^*)\|^2
\]
for any \(x \in \mathcal{C}\). Here, \(x^*\) is some fixed target parameter.

Our later analysis will be based on the assumption that \(L\) satisfies the RCG condition on to-be-specified sets of tensors with \(x^* = X^*\) being the true parameter tensor.

**Remark 3.1 (Interpretation of the RCG Condition)** The RCG condition is similar to the “regularity condition” appearing in recent nonconvex optimization literature (Chen and Candes, 2015; Candès, Li and Soltanolkotabi, 2015; Chi, Lu and Chen, 2019; Yonel and Yazici, 2020):
\[
(3.2) \quad \langle \nabla f(x), x - x^\# \rangle \geq \alpha \|x - x^\#\|^2 + \beta \|\nabla f(x)\|^2,
\]
where \(f(\cdot)\) is the objective function in their context and \(x^\#\) is the minimizer of \(f(\cdot)\). The RCG condition can be seen as a generalization of (3.2): in the deterministic case without statistical noise, the target \(x^*\) usually becomes an exact stationary point of \(f(\cdot)\) and (3.1) reduces to (3.2). In addition, it is worthy noting that RCG condition does not require the function \(f\) to be convex since \(x^*\) is only a fixed target parameter in the requirement (3.2) (also see Figure 1 in Chi, Lu and Chen (2019) for an example).

#### 3.2. Theoretical Analysis

We now consider a general setting that the loss function \(L\) satisfies the RCG condition in a constrained domain:
\[
(3.3) \quad \mathcal{C} = \{X \in \mathbb{R}^{p_1 \times p_2 \times p_3} : X = [S; U_1, U_2, U_3], U_k \in C_k, S \in C_S\},
\]
where the true parameter tensor \(X^*\) is feasible — that is, \(X^* \in \mathcal{C}\). Here, \(C_k\) and \(C_S\) are some convex and rotation invariant sets: for any \(U_k \in C_k\), \(S \in C_S\), we have \(U_k R_k \in C_k\) and \([S; R_1, R_2, R_3] \in C_S\) for arbitrary orthogonal matrices \(R_k \in \mathbb{O}_{r_k}\). Some specific problems of this general setting will be discussed in Section 4.

When \(L\) and \(X^*\) satisfy the condition above, we introduce the projected gradient descent in Algorithm 1. In addition to the vanilla gradient descent, the proposed Algorithm 1 includes
Algorithm 1 Projected Gradient Descent

Require: Initialization \(\left(\mathcal{S}(0), \mathbf{U}_1^{(0)}, \mathbf{U}_2^{(0)}, \mathbf{U}_3^{(0)}\right)\), constraint sets \(\{\mathcal{C}_k\}_{k=1}^3\), \(\mathcal{C}_S\), tuning parameters \(a, b > 0\), step size \(\eta\).

for all \(t = 0\) to \(T - 1\) do
  for all \(k = 1, 2, 3\) do
    \[
    \tilde{U}_k^{(t+1)} = \mathbf{U}_k^{(t)} - \eta \left(\nabla \mathbf{U}_k L(\mathcal{S}^{(t)}, \mathbf{U}_1^{(t)}, \mathbf{U}_2^{(t)}, \mathbf{U}_3^{(t)}) + a \mathbf{U}_k^{(t)} (\mathbf{U}_k^{(t)T} \mathbf{U}_k^{(t)} - b^2 \mathbf{I})\right)
    \]
    \[
    \mathbf{U}_k^{(t+1)} = \mathcal{P}_{\mathcal{C}_k}(\tilde{U}_k^{(t+1)}), \text{ where } \mathcal{P}_{\mathcal{C}_k}(\cdot) \text{ is the projection onto } \mathcal{C}_k.
    \]
  end for
  \[
  \mathcal{S}^{(t+1)} = \mathcal{S}^{(t)} - \eta \nabla \mathbf{S} L(\mathcal{S}^{(t)}, \mathbf{U}_1^{(t)}, \mathbf{U}_2^{(t)}, \mathbf{U}_3^{(t)})
  \]
  \[
  \mathcal{S}^{(t+1)} = \mathcal{P}_{\mathcal{C}_S}(\mathcal{S}^{(t+1)}), \text{ where } \mathcal{P}_{\mathcal{C}_S}(\cdot) \text{ is the projection onto } \mathcal{C}_S.
  \]
end for

return \(\mathcal{X}(T) = \mathcal{S}(T) \times_1 \mathbf{U}_1^{(T)} \times_2 \mathbf{U}_2^{(T)} \times_3 \mathbf{U}_3^{(T)}\)

multiple projection steps to ensure that \(\mathcal{X}^*\) is in the regularized domain \(\mathcal{C}\) throughout the iterations.

Suppose the true parameter \(\mathcal{X}^*\) is of Tucker rank-\((r_1, r_2, r_3)\). We also introduce the following value to quantify how different the \(\mathcal{X}^*\) is from being a stationary point of \(L(\mathcal{X}; \mathcal{D})\):

\[
\xi := \sup_{\mathcal{T} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, \|\mathcal{T}\|_F \leq 1} \|\langle \nabla L(\mathcal{X}^*), \mathcal{T} \rangle \|.
\]

Intuitively speaking, \(\xi\) measures the amplitude of \(\nabla L(\mathcal{X}^*)\) projected onto the manifold of rank-low tensors. In many statistical models, \(\xi\) essentially characterizes the amplitude of statistical noise. Specifically in the noiseless setting, \(\mathcal{X}^*\) is exactly a stationary point of \(L\), then \(\nabla L(\mathcal{X}^*) = 0\), \(\xi = 0\). In various probabilistic settings, a suitable \(L\) often satisfies \(\mathbb{E} \nabla L(\mathcal{X}^*) = 0\); then \(\xi\) reflects the reduction of variance of \(\nabla L(\mathcal{X}^*)\) after projection onto the low-rank tensor manifold. We also define

\[
\bar{\lambda} := \max \left\{\|\mathcal{M}_1(\mathcal{X}^*)\|, \|\mathcal{M}_2(\mathcal{X}^*)\|, \|\mathcal{M}_3(\mathcal{X}^*)\|\right\},
\]

\[
\lambda := \min \left\{\sigma_{r_1}(\mathcal{M}_1(\mathcal{X}^*)), \sigma_{r_2}(\mathcal{M}_2(\mathcal{X}^*)), \sigma_{r_3}(\mathcal{M}_3(\mathcal{X}^*))\right\},
\]

and \(\kappa = \bar{\lambda}/\lambda\) can be regarded as a tensor condition number, as similarly defined for matrices. It is not worth that the curvature of Tucker rank-\((r_1, r_2, r_3)\) tensor manifold on \(\mathcal{X}^*\) can be bounded by \(\lambda^{-1}\) (Lubich et al., 2013, Lemma 4.5).

We are now in position to establish a deterministic upper bound on the estimation error and a linear rate of convergence for the proposed Algorithm 1 when a warm initialization is provided. Specific initialization algorithms for different applications will be discussed in Section 4.

**Theorem 3.1 (Local Convergence)** Suppose \(L\) satisfies RCG(\(\alpha, \beta, \mathcal{C}\)) for \(\mathcal{C}\) defined in (3.3) and \(b \asymp \bar{\lambda}^{1/4}\), \(a = \frac{4a\bar{\lambda}^2}{3\kappa^2}\) in (2.4). Assume \(\mathcal{X}^* = \left[\mathcal{S}^*; \mathbf{U}_1^*, \mathbf{U}_2^*, \mathbf{U}_3^*\right]\) such that \(\mathbf{U}_k^T \mathbf{U}_k^* = b^2 \mathbf{I}_{r_k}, \mathbf{U}_k \in \mathcal{C}_k, k = 1, 2, 3, \) and \(\mathcal{S}^* \in \mathcal{C}_S\). Suppose the initialization \(\mathcal{X}(0) = \left[\mathcal{S}(0); \mathbf{U}_1^{(0)}, \mathbf{U}_2^{(0)}, \mathbf{U}_3^{(0)}\right]\) satisfies \(\|\mathcal{X}(0) - \mathcal{X}^*\|^2_F \leq \frac{\alpha \beta \kappa^2}{2}\) for some small constant \(c > 0\), \(\mathbf{U}_k^{(0)} \mathbf{U}_k^* = b^2 \mathbf{I}_{r_k}, \mathbf{U}_k \in \mathcal{C}_k, k = 1, 2, 3\) and \(\mathcal{S}(0) \in \mathcal{C}_S\). Also, the signal-noise-ratio satisfies \(\lambda^2 \geq C_0 \frac{\kappa^4}{\alpha \beta \kappa^2}\) for some universal constant \(C_0\). Then there exists a constant \(c > 0\) such that if \(\eta = \frac{\eta_0}{\sqrt{c}}\) for \(\eta_0 \leq c\), we have

\[
\|\mathcal{X}(t) - \mathcal{X}^*\|^2_F \leq C \left(\frac{\kappa^4}{\alpha \beta \kappa^2} \xi^2 + \kappa^2 \left(1 - \frac{2 \rho \alpha \beta \eta_0}{\kappa^2}\right)^t\right) \|\mathcal{X}(0) - \mathcal{X}^*\|^2_F.
\]
In addition, the following corollary provides a theoretical guarantee for the estimation loss of the proposed Algorithm 1 after a logarithmic number of iterations.

**Corollary 3.1** Suppose the conditions of Theorem 3.1 hold and $\alpha, \beta, \kappa$ are constants. Then after at most $T = \Omega\left(\log\left(\frac{\|\chi(0) - \chi^*\|_F}{\xi}\right)\right)$ iterations and for a constant $C$ that only relies on $\alpha, \beta, \kappa > 0$, we have

$$\|\chi(T) - \chi^*\|_F^2 \leq C \xi^2.$$  

**Remark 3.2** When $\nabla L(\chi^*) = 0$, i.e., there is no statistical noise or perturbation, we have $\xi = 0$. In this case, Theorem 3.1 and Corollary 3.1 imply that the proposed algorithm converges to the true target parameter $\chi^*$ at a linear rate:

$$\|\chi(t) - \chi^*\|_F^2 \leq C \left(1 - \frac{2\rho\kappa\beta\eta_0}{r^2}\right)^t \|\chi(0) - \chi^*\|_F^2.$$  

When $\nabla L(\chi^*) \neq 0$, we have $\xi > 0$ and $\chi^*$ is not an exact stationary point of the loss function $L$. Then the estimation error $\|\chi(t) - \chi^*\|_F$ is naturally not expected to go to zero, which matches the upper bounds of Theorem 3.1 and Corollary 3.1. In a statistical model where noise or perturbation is in presence, the upper bound on the estimation error can be determined by evaluating $\xi$ under the specific random environment and these bounds are often minimax-optimal. See Section 4 for more detail.

**Remark 3.3** If $C_S$ and $C_k$ are unbounded domains, then $C$ is the set of all rank-$(r_1, r_2, r_3)$ tensors, $P_{C_S}$, $P_{C_k}$ are identity operators, and the proposed Algorithm 1 essentially becomes the vanilla gradient descent. When $C_S$ and $C_k$ are non-trivial convex subsets, the projection steps ensure that $\chi(t) = [S(t); U_1^t, U_2^t, U_3^t] \in C$ and the RCG condition can be applied throughout the iterations. In fact, we found that the projection steps can be omitted in many numerical cases even if $L$ does not satisfy the RCG condition for the full set of low-rank tensors, such as the forthcoming Poisson and binomial tensor PCA. See Sections 4 and 7 for more discussions.

### 3.3. Proof Sketch of Main Results

We briefly discuss the idea for the proof of Theorem 3.1 here. The complete proof is provided in Appendix C. A key step in our analysis is to establish an error contraction inequality to characterize the estimation error of $\chi(t+1)$ based on the one of $\chi(t)$. Since the proposed non-convex gradient descent is performed on $S(t), U_1^t, U_2^t, U_3^t$ jointly in lieu of $\chi(t)$ directly, it becomes technically difficult to develop a direct link between $\|\chi(t+1) - \chi^*\|_F^2$ and $\|\chi(t) - \chi^*\|_F^2$. To overcome this difficulty, a “lifting” scheme was proposed and widely used in the recent literature on low-rank asymmetric matrix optimization (Tu et al., 2016; Zhu et al., 2017; Park et al., 2018): one can factorize any rank-$r$ matrix estimator $A(t)$ and the target matrix parameter $A^*$ into $A(t) = U(t)(V(t))^T, A^* = U^*(V^*)^T$, where $U(t), V(t)$ (or $U^*, V^*$) both have $r$ columns and share the same singular values. Then, one can stack them into one matrix

$$W(t) = \begin{bmatrix} U(t) \\ V(t) \end{bmatrix}, \quad W^* = \begin{bmatrix} U^* \\ V^* \end{bmatrix}.$$  

By establishing the equivalence between $\min_{R \in \mathbb{R}^r} \|W(t) - W^*R\|_F^2$ and $\|A(t) - A^*\|_F^2$, and analyzing on $\min_{R \in \mathbb{R}^r} \|W(t) - W^*R\|_F^2$, a local convergence of $A(t)$ to $A^*$ can be established. However, the “lifting” scheme is not applicable to the tensor problem here since $U_1, U_2, U_3, S$ have distinct shapes and cannot be simply stacked together. To overcome
this technical issue in the generalized tensor estimation problems, we propose to assess the following criterion:

\[
E(t) = \min_{R_k \in O(p_k, r_k)} \left\{ \sum_{k=1}^{3} \left\| U_k(t) - U_k^* R_k \right\|_F^2 + \left\| S(t) - [S^*; R_1^T, R_2^T, R_3^T] \right\|_F^2 \right\}.
\]

Intuitively, \(E(t)\) measures the difference between a pair of tensor components \((S^*, U_1, U_2, U_3)\) and \((S(t), U_1(t), U_2(t), U_3(t))\) under rotation. The introduction of \(E(t)\) enables a convenient error contraction analysis as being an additive form of tensor components. In particular, the following lemma exhibits that \(E(t)\) is equivalent to the estimation error \(\|X(t) - X^*\|_F^2\) under regularity conditions.

**Lemma 3.1 (An informal version of Lemma E.2)**

\[
cE(t) \leq \|X(t) - X^*\|_F^2 + C \sum_{k=1}^{3} \left\| (U_k(t))^T U_k^* - U_k^* U_k^T \right\|_F^2 \leq CE(t)
\]

under the regularity conditions to be specified in Lemma E.2.

Note that there is no equivalence between \(E(t)\) and \(\|X(t) - X^*\|_F^2\) unless we force \(U_k\) and \(U_k^*\) have similar singular structures, and this is the reason why we introduce the regularizer term in (2.4) to keep \(U_k^*\) from being singular.

Based on Lemma 3.1, the proof of Theorem 3.1 reduces to establishing an error contraction inequality between \(E(t)\) and \(E(t+1)\):

\[
E(t+1) \leq (1 - \gamma) E(t) + C\xi^2
\]

for constants \(0 < \gamma < 1\) and \(C > 0\). Define the best rotation matrices

\[
(R_1(t), R_2(t), R_3(t)) = \arg \min_{R_k \in O(p_k, r_k)} \left\{ \sum_{k=1}^{3} \left\| U_k(t) - U_k^* R_k \right\|_F^2 + \left\| S(t) - [S^*; R_1^T, R_2^T, R_3^T] \right\|_F^2 \right\}.
\]

By plugging in the gradient of \(L(X)\) and \(X = S \times_1 U_1 \times_2 U_2 \times_3 U_3\), we can show

\[
\left\| U_k(t+1) - U_k^* R_k(t+1) \right\|_F^2 - \left\| U_k(t) - U_k^* R_k(t) \right\|_F^2 
\approx -2\eta \left\langle X(t), \nabla L(X(t)) \right\rangle - \frac{\eta y}{2} \left\| U_k(t) U_k^T - b^2 I_{p_k} \right\|_F^2,
\]

\[
\left\| S(t+1) - [S^*; R_1(t+1)^T, R_2(t+1)^T, R_3(t+1)^T] \right\|_F^2 - \left\| S(t) - [S^*; R_1(t)^T, R_2(t)^T, R_3(t)^T] \right\|_F^2
\approx -2\eta \left\langle X(t), \nabla L(X(t)) \right\rangle,
\]

where

\[X_k(t) := S(t) \times_k (U_k R_k(t)) \times_{k+1} U_{k+1}(t) \times_{k+2} U_{k+2}(t), \quad k = 1, 2, 3;\]

and \[X_S(t) = [S^*; U_1(t) R_1(t)^T, U_2(t) R_2(t)^T, U_3(t) R_3(t)^T].\]

Note \(E(t+1) - E(t)\) corresponds to the summation of (3.7) and (3.8), whose right hand sides are dominated by the inner product between \(\nabla L(X(t))\) and \(3X(t) - \sum_{k=1}^{3} X_k(t) - X_S(t)\). We develop a new tensor perturbation Lemma to characterize \(3X(t) - \sum_{k=1}^{3} X_k(t) - X_S(t)\).
Lemma 3.2 (An informal version of Lemma E.3) Under regularity conditions to be specified in Lemma E.3, we have

\[ \mathcal{X}^{(t)} - \mathcal{X}^* = \left( \mathcal{X}^{(t)} - \mathcal{X}^{(t)}_S \right) + \sum_{k=1}^{3} \left( \mathcal{X}^{(t)} - \mathcal{X}^{(t)}_k \right) + \mathcal{H}_\varepsilon, \]

where \( \mathcal{H}_\varepsilon \) is some low-rank residual tensor with \( \| \mathcal{H}_\varepsilon \|_F^2 = o \left( (E^{(t)})^2 \right) \).

Combining (3.7)(3.8) and Lemma 3.2, we can connect \( E^{(t+1)} \) and \( E^{(t)} \) as

\[ E^{(t+1)} \approx E^{(t)} - 2\eta \left( \mathcal{X}^{(t)} - \mathcal{X}^* - \mathcal{H}_\varepsilon, \nabla \mathcal{L} \left( \mathcal{X}^{(t)} \right) \right) - \frac{\alpha \eta}{2} \sum_{k=1}^{3} \left\| U_k^{(t)\top} U_k^{(t)} - U_k^{*\top} U_k^* \right\|^2_F. \]

Then, we introduce another decomposition

\[ \left( \mathcal{X}^{(t)} - \mathcal{X}^* - \mathcal{H}_\varepsilon, \nabla \mathcal{L} \left( \mathcal{X}^{(t)} \right) \right) = \underbrace{\left( \mathcal{X}^{(t)} - \mathcal{X}^*, \nabla \mathcal{L} \left( \mathcal{X}^{(t)} \right) - \nabla \mathcal{L} \left( \mathcal{X}^* \right) \right)}_{A_1} \]

\[ - \underbrace{\left( \mathcal{H}_\varepsilon, \nabla \mathcal{L} \left( \mathcal{X}^{(t)} \right) - \nabla \mathcal{L} \left( \mathcal{X}^* \right) \right)}_{A_2} + \underbrace{\left( \mathcal{X}^{(t)} - \mathcal{X}^* + \mathcal{H}_\varepsilon, \nabla \mathcal{L} \left( \mathcal{X}^* \right) \right)}_{A_3}. \]

The three terms can be bounded separately:

\[ A_1 \geq \alpha \| \mathcal{X}^{(t)} - \mathcal{X}^* \|_F^2 + \beta \left\| \nabla \mathcal{L} \left( \mathcal{X}^{(t)} \right) - \nabla \mathcal{L} \left( \mathcal{X}^* \right) \right\|^2_F, \]

\[ |A_2| \leq \frac{\beta}{2} \left\| \nabla \mathcal{L} \left( \mathcal{X}^{(t)} \right) - \nabla \mathcal{L} \left( \mathcal{X}^* \right) \right\|^2_F + \frac{1}{2\beta} \| \mathcal{H}_\varepsilon \|_F^2, \]

\[ |A_3| \leq \sqrt{E^{(t)}} \cdot \xi \leq cE^{(t)} + C\xi^2. \]

Here the first inequality comes from RCG condition; the second inequality comes from Cauchy-Schwarz inequality and the fact \( ab \leq \frac{1}{2} (a^2 + b^2) \); and the last inequality utilizes the definition of \( \xi \), as well as Lemma 3.2. Combining (3.11) and (3.12), we obtain

\[ \left( \mathcal{X}^{(t)} - \mathcal{X}^* - \mathcal{H}_\varepsilon, \nabla \mathcal{L} \left( \mathcal{X}^{(t)} \right) \right) + \left( \mathcal{X}^{(t)} - \mathcal{X}^* + \mathcal{H}_\varepsilon, \nabla \mathcal{L} \left( \mathcal{X}^* \right) \right) \geq \left( \alpha \| \mathcal{X}^{(t)} - \mathcal{X}^* \|_F^2 + \beta \left\| \nabla \mathcal{L} \left( \mathcal{X}^{(t)} \right) - \nabla \mathcal{L} \left( \mathcal{X}^* \right) \right\|^2_F \right) \]

\[ - \frac{\beta}{2} \left\| \nabla \mathcal{L} \left( \mathcal{X}^{(t)} \right) - \nabla \mathcal{L} \left( \mathcal{X}^* \right) \right\|^2_F - \left( cE^{(t)} + C\xi^2 \right). \]

Then by choosing a suitable step size \( \eta \) and applying (3.10) together with (3.13), one obtains

\[ E^{(t+1)} \leq E^{(t)} + cE^{(t)} + C\xi^2 - c_2 \left\| \mathcal{X}^{(t)} - \mathcal{X}^* \right\|^2_F - c_3 \sum_{k=1}^{3} \left\| U_k^{(t)\top} U_k^{(t)} - U_k^{*\top} U_k^* \right\|^2_F. \]

Applying the equivalence between \( \| \mathcal{X}^{(t)} - \mathcal{X}^* \|_F^2 + C \sum_k \left\| U_k^{(t)\top} U_k^{(t)} - U_k^{*\top} U_k^* \right\|^2_F \) and \( E^{(t)} \) (Lemma 3.1), we can obtain (3.6) and finish the proof of Theorem 3.1.

4. Applications of Generalized Tensor Estimation. Next, we apply the deterministic result to a number of generalized tensor estimation problems, including sub-Gaussian tensor PCA, tensor regression, Poisson tensor PCA, and binomial tensor PCA to obtain the estimation error bound of (projected) gradient descent. In each case, Algorithm 1 is used with different initialization schemes specified by the problem settings. All the proofs are provided in Appendix D. In addition, the generalized tensor estimation framework covers many other problems. A non-exhaustive list is provided in the introduction. See Section 8 for more discussions.
4.1. Sub-Gaussian Tensor PCA. Suppose we observe $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, where $E\mathcal{Y} = \mathcal{X}^*$, $\mathcal{X}^*$ is Tucker low-rank, and $\{Y_{ijk} - \mathcal{X}^*_{ijk}\}_{ijk}$ are independent and sub-Gaussian distributed. In literature, much attention has been focused on various setups related to this model, e.g., $Y_{ijk} - \mathcal{X}^*_{ijk}$ are i.i.d. Gaussian, $\mathcal{X}^*$ is sparse, symmetric, rank-1, or CP-low-rank, etc (Richard and Montanari, 2014; Sun and Luo, 2015; Perry, Wein and Bandeira, 2016; Montanari, Reichenman and Zeitouni, 2017; Lesieur et al., 2017; Zhang and Han, 2018; Chen, 2019). Particularly when $\{Y_{ijk} - \mathcal{X}^*_{ijk}\}_{ijk}$ are i.i.d. Gaussian distributed, it has been shown that the higher-order orthogonal iteration (HOOI) (De Lathauwer, De Moor and Vandewalle, 2000b) achieves the optimal statistical performance on the estimation of $\mathcal{X}^*$ (Zhang and Xia, 2018). It is however unclear whether HOOI works in the more general heteroskedastic setting, where the entries of $\mathcal{Y}$ have different variances.

Departing from the existing methods, we consider the estimation of $\mathcal{X}^*$ via minimizing the quadratic loss function $L(\mathcal{X}) = \frac{1}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2$ using gradient descent. It is easy to check that $L$ satisfies $RCG(\frac{1}{2}, \frac{1}{2}, \mathbb{R}^{p_1 \times p_2 \times p_3})$, so the projection steps in Algorithm 1 can be skipped throughout the iterations. To accommodate possible heteroskedastic noise, we apply HeteroPCA (Zhang, Cai and Wu, 2018), an iterative algorithm for PCA when heteroskedastic noise appears instead of the regular PCA for initialization. (The implementation of HeteroPCA in Algorithm 2 is provided in Appendix A).

**Algorithm 2 Initialization for Sub-Gaussian Tensor PCA**

**Require:** $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, Tucker rank $(r_1, r_2, r_3)$, scaling parameter $b$.

$\mathcal{X} = \mathcal{Y}$, $\tilde{U}_k = \text{HeteroPCA}_{\mathcal{X}} \left(\mathcal{M}_k(\mathcal{X})\mathcal{M}_k(\mathcal{X})^\top\right)$ for $k = 1, 2, 3$

$\tilde{S} = [\tilde{S}; \tilde{U}^1_1, \tilde{U}^2_2, \tilde{U}^3_3]$

$S^{(0)} = \tilde{S}/b^3$, $U^{(0)}_k = b\tilde{U}^{(0)}_k$ for $k = 1, 2, 3$

return $(S^{(0)}, U^{(0)}_1, U^{(0)}_2, U^{(0)}_3)$

Now we can establish the theoretical guarantee of Algorithms 1 and 2 for sub-Gaussian tensor PCA based on the deterministic result in Theorem 3.1.

**Theorem 4.1** Suppose we observe $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, where $E\mathcal{Y} = \mathcal{X}^* = [S; U_1, U_2, U_3]$ is Tucker rank-$(r_1, r_2, r_3)$, $U_k \in \mathbb{O}_{p_1 r_k}$ and $\|U_k\|_{2,\infty} \leq c$ for some constant $c > 0$. Suppose all entries of $\mathcal{Y} - \mathcal{X}^*$ are independent mean-zero sub-Gaussian random variables such that $\|Y_{ijk} - \mathcal{X}^*_{ijk}\|_{\psi_2} = \sup_{q \geq 1}E(\|Z_{ijk}\|^q)^{1/q} / q^{1/2} \leq \sigma$.

Assume $\lambda \sigma \geq C_1 \bar{p}^{3/4} \bar{r}^{1/4}$. Then with probability at least $1 - \exp(\epsilon\bar{p})$, Algorithms 1 and 2 yield

$$\|\mathcal{X} - \mathcal{X}^*\|_F^2 \leq C_2 \sigma^2 \left( r_1 r_2 r_3 + \sum_{k=1}^{3} p_k r_k \right),$$

where $C_1, C_2$ are constants that do not depend on $p_k$ and $r_k$.

**Remark 4.1** The proposed method turns out to achieve the minimax optimal rate of estimation error in a general class of sub-Gaussian tensor PCA settings since the order of upper bound (4.1) matches the lower bound in literature (Zhang and Xia, 2018, Theorem 3). Moreover, the condition $\lambda \sigma \geq C_1 \bar{p}^{3/4} \bar{r}^{1/4}$ is optimal w.r.t. $\bar{p}$ in the sense that all the polynomial-time feasible algorithms cannot achieve consistent estimation when $\lambda \sigma < \bar{p}^{3/4 - \epsilon}$ for any $\epsilon > 0$ (Zhang and Xia, 2018, Theorem 4).
4.2. Low-rank Tensor Regression. Motivated by applications of neuroimaging analysis (Zhou, Li and Zhu, 2013; Li and Zhang, 2017; Guhaniyogi, Qamar and Dunson, 2017), spatio-temporal forecasting (Bahadori, Yu and Liu, 2014), high-order interaction pursuit (Hao, Zhang and Cheng, 2019), longitudinal relational data analysis (Hoff, 2015), 3D imaging processing (Guo, Kotsia and Patras, 2012), among many others, we consider the low-rank tensor regression next. Suppose we observe a collection of data \( D = \{y_i, A_i\}_{i=1}^n \) that are associated through the following equation:

\[
y_i = \langle A_i, X^* \rangle + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n.
\]

By exploiting the negative log-likelihood, it is natural to set \( L \) to be the squared loss function

\[
L(X; D) = \sum_{i=1}^n (\langle A_i, X \rangle - y_i)^2.
\]

To estimate \( X^* \), we first perform spectral method (Algorithm 3) to obtain initializer \( X^{(0)} = [S^{(0)}, U_1^{(0)}, U_2^{(0)}, U_3^{(0)}] \), then perform the gradient descent (Algorithm 1) without the projection steps to obtain the final estimator \( \hat{X} \). A key step of Algorithm 3 is HOSVD or HOOI, which are described in detail in Appendix A.

**Algorithm 3 Initialization of Low-rank Tensor Regression**

**Require:** \( \{A_i, y_i\}, i = 1, \ldots, n, \) rank \( (r_1, r_2, r_3) \), scaling parameter \( h \).

\[
\hat{X} = \frac{1}{n} \sum_{i=1}^n y_i A_i
\]

\((\tilde{S}, \tilde{U}_1, \tilde{U}_2, \tilde{U}_3) = \text{HOSVD} \left( \hat{X} \right) \) or \((\tilde{S}, \tilde{U}_1, \tilde{U}_2, \tilde{U}_3) = \text{HOOI} \left( \hat{X} \right) \)

\[
U_k^{(0)} = \tilde{U}_k, \quad \text{for } k = 1, 2, 3
\]

\[
S^{(0)} = \tilde{S}/h^3
\]

**return** \( (S^{(0)}, U_1^{(0)}, U_2^{(0)}, U_3^{(0)}) \)

For technical convenience, we assume the covariates \( \{A_i\}_{i=1}^n \) are randomly designed that all entries of \( A_i \) are i.i.d. drawn from sub-Gaussian distribution with mean 0 and variance 1. The following theorem gives an estimation error upper bound for Algorithms 1 and 3.

**Theorem 4.2** Consider the low-rank tensor regression model (4.2). Suppose \( \sigma^2 \leq C_1 \|X^*\|_F^2 \), \( \lambda \geq C_2 \), and the sample size \( n \geq C_3 \max \{p_0^{3/2}, p_0^{2/2}, p_2, p_4^2, p_3^{1/2} \} \) for constants \( C_1, C_2, C_3 > 0 \). Then with probability at least \( 1 - \exp \left\{ -c (r_1 r_2 r_3 + \sum_{k=1}^3 p_k r_k) \right\} \), the output of Algorithms 1 and 3 satisfies

\[
\|\hat{X} - X^*\|_F^2 \leq C_1 \sigma^2 \left( r_1 r_2 r_3 + \sum_{k=1}^3 p_k r_k \right) / n,
\]

where \( C_1, C_2, C_3, C \) are constants depending only on \( \kappa \) and \( c > 0 \) is a universal constant.

Theorem 4.2 together with the lower bound in (Zhang et al., 2019, Theorem 5) shows that the proposed procedure achieves the minimax optimal rate of estimation error in the class of all \( p_1 \)-by-\( p_2 \)-by-\( p_3 \) tensors with rank-(\( r_1, r_2, r_3 \)) for tensor regression.

**Remark 4.2** The assumption on the covariates \( \{A_i\}_{i=1}^n \) in Theorem 4.2 ensures that \( \hat{X} \) in Algorithm 3 is an unbiased estimator of \( X^* \). Such a setting has been considered as a benchmark setting in the high-dimensional statistical inference literature (see, e.g., Candes and Plan (2011); Chen, Raskutti and Yuan (2019); Javanmard et al. (2018)). When \( \{A_i\}_{i=1}^n \) are
We thus introduce the following assumption on the Poisson likelihood function is not strongly convex and smooth in the unbounded domain.

\[
\hat{X}' = \arg \min_{X \in \mathbb{R}^{p_1 \times p_2 \times p_3}} \sum_{k=1}^{3} \| M_k(X) \|_*, \quad \text{s.t. } A(X) = y.
\]

(4.3) is equivalent to a semidefinite programming and can be solved by the interior-point method (Gandy, Recht and Yamada, 2011).

**Remark 4.3** There is a significant gap between the required sample size in Theorem 4.2 \((O(p^{3/2}r))\) and the possible sample size lower bound, i.e., the degree of freedom of all rank-\(r\) dimension-\(p\) tensor \((O(r^3 + pr))\). The existing algorithms achieving the sample size lower bound are often NP-hard to compute and thus intractable in practice. We also note that the existence of a tractable algorithm for tensor completion that provably works with less than \(p^{3/2-\varepsilon}\) measurements would disapprove an open conjecture in theoretical computer science on strongly random 3-SAT (Barak and Moitra, 2016, Corollary 16). Since tensor completion can be seen as a special case of tensor recovery, this suggests that it may be impossible to substantially improve the sample complexity required in Theorem 4.2 using a polynomial-time algorithm. Therefore, our procedure can be taken as the first computationally efficient algorithm to achieve minimax optimal rate of convergence and exact recovery in the noiseless setting as illustrated in Table 1.

4.3. Poisson Tensor PCA. Tensor data with count values commonly arise from various scientific applications, such as the photon-limited imaging (Timmerman and Nowak, 1999; Willett and Nowak, 2007; Salmon et al., 2014), online click-through data analysis (Shan et al., 2016; Sun and Li, 2016), and metagenomic sequencing (Flores et al., 2014). In this section, we consider the Poisson tensor PCA model: assume we observe \(Y \in \mathbb{N}^{p_1 \times p_2 \times p_3}\) such that

\[
Y_{ijk} \sim \text{Poisson}(I \exp(X^*_{ijk})) \quad \text{independently},
\]

where \(X^*\) is the low-rank tensor parameter and \(I > 0\) is the intensity parameter. When \(X^*\) is entry-wise bounded (Assumption 4.1), one can set \(I\) as the average intensity of all entries of \(Y\) so that \(I\) essentially quantifies the signal-to-noise ratio. Rather than estimating \(I \exp(X^*)\), we focus on estimating \(X^*\), the key tensor that captures the salient geometry or structure of the data. Then, the following negative log-likelihood is a natural choice of the loss function for estimating \(X^*\),

\[
L(X) = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \sum_{k=1}^{p_3} (-Y_{ijk}X_{ijk} + I \exp(X_{ijk})).
\]

Unfortunately, \(L(X)\) defined in (4.5) satisfies RCG(\(\alpha, \beta, C\)) only for a bounded set \(C\) since the Poisson likelihood function is not strongly convex and smooth in the unbounded domain. We thus introduce the following assumption on \(X^*\) to ensure that \(X^*\) is in a bounded set \(C\).

**Assumption 4.1** Suppose \(X^* = [S^*; U^*_1, U^*_2, U^*_3]\), where \(U^*_k \in \mathcal{O}_{p_k} \) is a \(p_k\)-by-\(r_k\) orthogonal matrix for \(k = 1, 2, 3\). There exist some constants \(\{\mu_k\}_{k=1}^{3}\) such that \(\| U^*_k \|_{p,\infty} \leq \frac{p_k^{1/2}}{r_k} \) for \(k = 1, 2, 3\) and \(\overline{\lambda} \leq B \sqrt{\frac{\| H^*_1 \|_{2,\infty}}{\| H^*_3 \|_{2,\infty} \mu_k p_k}}\) where \(\overline{\lambda} := \max_k \| M_k(S^*) \|\). Here, \(\| U^*_k \|_{2,\infty} = \max_i \| (U^*_k)_i \|_2\) is the largest row-wise \(\ell_2\) norm of \(U^*_k\).

Assumption 4.1 requires that the loading \(U^*_k\) satisfies the incoherence condition, i.e., the amplitude of the tensor is “balanced” in all parts. Previously, the incoherence condition and its
variations were commonly used in the matrix estimation literature (Candès and Recht, 2009; Ma and Ma, 2017) and Poisson-type inverse problems (e.g., Poisson sparse regression (Jiang, Raskutti and Willett, 2015, Assumption 2.1), Poisson matrix completion (Cao and Xie, 2015, Assumption 2.1), Poisson matrix completion (Cao and Xie, 2015, Equation (7)), compositional matrix estimation (Cao, Zhang and Li, 2019, Equation (7)), Poisson auto-regressive models (Hall, Raskutti and Willett, 2016)). Assumption 4.1 also requires an upper bound on the spectral norm of each matricization of the core tensor $S^*$. Together with the incoherence condition on $U_k^*$, this condition guarantees that $X^*$ is entry-wise upper bounded by $B$. In fact, the entry-wise bounded assumption is also widely used in high-dimensional matrix/tensor generalized linear models since it guarantees the local strong convexity and smoothness of the negative log-likelihood function (Ma and Ma, 2017; Wang and Li, 2018; Xu, Hu and Wang, 2019).

Next, we set $(\{C_k\}_{k=1}^3, C_S)$ as follows:

$$C_k = \left\{ U_k \in \mathbb{R}^{p_k \times r_k} : \|U_k\|_{2,\infty} \leq b \sqrt{\frac{\mu_k r_k}{p_k}} \right\},$$

(4.6)

$$C_S = \left\{ S \in \mathbb{R}^{r_1 \times r_2 \times r_3} : \max_k \|M_k(S)\| \leq b^{-3} B \sqrt{\frac{\Pi_{k=1}^3 p_k}{\Pi_{k=1}^3 \mu_k r_k}} \right\}.$$

Specifically for the Poisson tensor PCA, we can prove that if Assumption 4.1 holds, the loss function (4.5) satisfies $RCG(\alpha, \beta, \{C_k\}_{k=1}^3, C_S)$ for constants $\alpha, \beta$ that only depend on $I$ and $B$ (see the proof of Theorem 4.3 for details). We can also show that the following Algorithm 4 provides a sufficiently good initialization with high probability.

**Algorithm 4 Initialization for Poisson Tensor PCA**

**Require:** Initialization observation tensor $X \in \mathbb{N}^{[p_1 \times p_2 \times p_3]}$, Tucker rank $(r_1, r_2, r_3)$, scaling parameter $b$, intensity parameter $I$.

$\hat{X} = \log \left( (Y_{jkl} + \frac{1}{2}) / I \right)$

$(\hat{S}, \hat{U}_1, \hat{U}_2, \hat{U}_3) = \text{HOSVD} \left( \hat{X} \right)$ or $(\hat{S}, \hat{U}_1, \hat{U}_2, \hat{U}_3) = \text{HOOI} \left( \hat{X} \right)$

$U_k^{(0)} = b \hat{U}_k$, for $k = 1, 2, 3$

$S^{(0)} = \hat{S} / b^3$

**return** $(S^{(0)}, U_1^{(0)}, U_2^{(0)}, U_3^{(0)})$

Now we establish the estimation error upper bound for Algorithms 1 and 4.

**Theorem 4.3** Suppose Assumption 4.1 holds and $I > C_1 \max \{ \bar{p}, \lambda^{-2} \sum_{k=1}^3 (p_{-k} r_k + p_k r_k) \}$, where $p_{-k} := p_1 p_2 p_3 / p_k$. Then with probability at least $1 - c / (p_1 p_2 p_3)$, the output of Algorithms 1 and 4 yields

$$\|\hat{X} - X^*\|_F^2 \leq C_2 I^{-1} \left( r_1 p_2 r_3 + \sum_{k=1}^3 p_k r_k \right).$$

Here $C_1, C_2$ are constants that do not depend on $p_k$ or $r_k$.

We further consider the following class of low-rank tensors $F_{p,r}$, where the restrictions in $F_{p,r}$ correspond to the conditions in Theorem 4.3:

$$F_{p,r} = \left\{ X = [S; U_1, U_2, U_3] : \max_k \|M_k(S)\| \leq B \sqrt{\frac{\Pi_{k=1}^3 p_k}{\Pi_{k=1}^3 \mu_k r_k}} \right\}.$$

(4.7)

With some technical conditions on tensor rank and the intensity parameter, we can develop the following lower bound in estimation error for Poisson PCA.
Theorem 4.4 (Lower Bound for Poisson tensor PCA) \textit{Assume } $\bar{r} \leq C_1 p^{1/2}$, $\bar{r} > C_2$ and $\min_k \mu_k \geq C_3$ for constants $C_1, C_2, C_3 > 1$. Suppose one observes $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, where $\mathcal{Y}_{jkl} \sim \text{Poisson} \left( I \exp(\lambda_{jkl}) \right)$ independently, $\lambda \in \mathcal{F}_{p,r}$, and $I \geq c_0$. There exists a uniform constant $c$ that does not depend on $p_k$ or $r_k$, such that

$$\inf \sup_{\hat{X}} \mathbb{E} \left\| \hat{X} - X \right\|_F^2 \geq c I^{-1} \left( r_1 r_2 r_3 + \sum_{k=1}^{3} p_k r_k \right).$$

Theorems 4.3 and 4.4 together yield the optimal rate of estimation error for Poisson tensor PCA problem over the class of $\mathcal{F}_{p,r}$:

$$\inf \sup_{\hat{X}} \mathbb{E} \left\| \hat{X} - X \right\|_F^2 \asymp I^{-1} \left( r_1 r_2 r_3 + \sum_{k=1}^{3} p_k r_k \right).$$

4.4. Binomial Tensor PCA. The binomial tensor data commonly arise in the analysis of proportion when raw counts are available. For example, in the Human Mortality Database (Wilmoth and Shkolnikov, 2006), the number of deaths and the total number of population are summarized into a three-way tensor, where the $x$-, $y$-, $z$-coordinates are counties, ages, and years, respectively. Given the sufficiently large number of population in each country, one can generally assume that each entry of this data tensor satisfies the binomial distribution independently.

Suppose we observe a count tensor $\mathcal{Y} \in \mathbb{N}^{p_1 \times p_2 \times p_3}$ and a total population tensor $\mathcal{N} \in \mathbb{N}^{p_1 \times p_2 \times p_3}$ such that $\mathcal{Y}_{jkl} \sim \text{Binomial}(N_{jkl}, p_{jkl})$ independently. Here, $p_{jkl} \in [0, 1]^{p_1 \times p_2 \times p_3}$ is a probability tensor linked to an underlying latent parameter $\lambda^* \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ through $p_{jkl} = s(\lambda_{jkl})$, where $s(x) = 1/(1 + e^{-x})$ is the sigmoid function. Our goal is to estimate $\lambda^*$. To this end, we consider to minimize the following loss function:

$$L(\lambda) = - \sum_{jkl} \left( \hat{P}_{jkl} \mathcal{Y}_{jkl} + \log (1 - \sigma(\lambda_{jkl})) \right),$$

where $\hat{P}_{jkl} := \mathcal{Y}_{jkl}/\mathcal{N}_{jkl}$.

We assume $\lambda^*$ satisfies Assumption 4.1 for the same reasons as in Poisson tensor PCA. We propose to estimate $p^*$ by applying Algorithm 5 (initialization) and Algorithm 1 (projected gradient descent) with the following constraint sets:

$$C_k = \left\{ U_k \in \mathbb{R}^{p_k \times r_k} : \| U_k \|_{2,\infty} \leq b \sqrt{\mu_k r_k / p_k} \right\},$$

$$C_S = \left\{ S \in \mathbb{R}^{r_1 \times r_2 \times r_3} : \max_k \| \mathcal{M}_k(S) \| \leq b^3 B \sqrt{\Pi_{k=1}^{3} p_k} \right\}.$$ 

We have the following theoretical guarantee for the estimator obtained by Algorithms 1 and 5 in binomial tensor PCA.

Theorem 4.5 (Upper Bound for Binomial Tensor PCA) Suppose Assumption 4.1 is satisfied and $N = \min_{jkl} N_{jkl}$ satisfies $N \geq C_1 \max \left\{ \Pi_3 p_k^2 \lambda_{jkl}^{-2} \sum_k (p_{-k} r_k + p_k r_k) \right\}$. Then with probability at least $1 - c/(p_1 p_2 p_3)$, we have the following estimation upper bound for the output of Algorithms 1 and 5:

$$\left\| \hat{X} - \lambda^* \right\|_F^2 \leq C_2 N^{-1} \left( r_1 r_2 r_3 + \sum_{k=1}^{3} p_k r_k \right).$$

Here, $C_1, C_2$ are some absolute constants that do not depend on $p_k$ or $r_k$. 
Let $\mathcal{F}_{p,r}$ be the class of low-rank tensors defined in (4.7). We can prove the following lower bound result, which establishes the minimax optimality of the proposed procedure over the class of $\mathcal{F}_{p,r}$ in binomial tensor PCA.

**Theorem 4.6 (Lower Bound for Binomial Tensor PCA)** Denote $N = \min_{jkl} N_{jkl}$. Assume $\tau \leq C_1 p^{1/2}$, $\delta > C_2$ and $\min_{jkl} \mu k \geq C_3$ for some constants $C_1, C_2, C_3 > 1$. Suppose one observes $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, where $\mathcal{Y}_{jkl} \sim \text{Binomial} (N_{jkl}, \sigma(X_{jkl}))$ independently, $X^* \in \mathcal{F}_{p,r}$, and $\max_{jkl} N_{jkl} \leq C \min_{jkl} N_{jkl}$. There exists constant $c$ that does not depend on $p_k$ or $r_k$, such that

$$\inf_{\hat{X}} \sup_{X \in \mathcal{F}_{p,r}} \| \hat{X} - X \|_F^2 \geq c N^{-1} \left( r_1 r_2 r_3 + \sum_{k=1}^3 p_k r_k \right).$$

5. **Rank Selection.** The tensor rank $(r_1, r_2, r_3)$ is required as an input to Algorithm 1 and plays a crucial role in the proposed non-convex optimization framework. While various empirical methods have been proposed for rank selection in specific applications of low-rank tensor estimation (e.g., Yokota, Lee and Cichocki (2016)), there is a paucity of theoretical guarantees in the literature. In this section, we provide a rank estimation procedure with provable guarantees. Recall that in each application in Section 4, we first specify the initialization $X^{(0)} = [S^{(0)}, U_1^{(0)}, U_2^{(0)}, U_3^{(0)}]$ based on a spectral algorithm on some preliminary tensor $\tilde{X}$ (see their definitions in Algorithms 2-5). Since $\tilde{X}$ reflects the target tensor $X^*$ and $\sigma_s(\mathcal{M}_k(X^*)) = 0$ for $s \geq r_k + 1$, we consider the following rank selection method by exploiting the singular values of $\tilde{X}$:

$$\hat{r}_k = \max \left\{ r : \sigma_r \left( \mathcal{M}_k(\tilde{X}) \right) \geq t_k \right\}, \quad k = 1, 2, 3.$$

Here, $t_k > 0$ is the thresholding level whose value depends on specific problem settings. Next, we specifically consider the sub-Gaussian tensor PCA and tensor regression.

**Proposition 5.1** Suppose $\text{rank}(X^*) = (r_1, r_2, r_3)$, $r_k = o(p_k)$, $p_k = o(p_{-k})$. Define $\hat{r}_k$ as (5.1) and

$$\delta_k := \text{Median} \left\{ \sigma_1(\mathcal{M}_k(\tilde{X})), \ldots, \sigma_{p_k}(\mathcal{M}_k(\tilde{X})) \right\}.$$

(a) In sub-Gaussian tensor PCA (Section 4.1), we set $\tilde{X} = \mathcal{Y}$, $t_k = 1.5 \delta_k$. Suppose $\lambda \geq C \bar{\sigma}$. Then, we have $\hat{r}_k = r_k, \forall k \in [d]$ with probability at least $1 - C e^{-\bar{\sigma}^2}$. 

**Remark 4.4** We assume $N \geq C_1 \max \left\{ \bar{p}, \lambda^{-2} \sum_k (p_{-k} r_k + p_k r_k) \right\}$ in Theorem 4.5 as a technical condition to prove the estimation error upper bound of $\hat{X}$. $N$ here essentially characterizes the signal-noise ratio of the binomial tensor PCA problem.
(b) In low-rank tensor regression (Section 4.2), we set \( \hat{X} = \frac{1}{n} \sum_{i=1}^{n} y_i A_i, t_k = 1.5 \delta_k \). Suppose \( n \geq C K^2 \rho^3/2 \log \hat{r} \) and \( \Delta \geq C \frac{\rho r}{\sqrt{n}} \). Then, \( \hat{r}_k = r_k, \forall k \in [d] \) with probability at least \( 1 - Ce^{-c \rho} \).

In practice, we can also apply a simple criterion of the cumulative percentage of total variation (Jolliffe, 1986, Chapter 6.1.1) originating from principle component analysis:

\[
\tilde{r}_k = \arg \min \left\{ r : \sum_{i=1}^{r} \sigma_i^2(\mathcal{M}_k(\tilde{X})) / \sum_{i=1}^{p_k} \sigma_i^2(\mathcal{M}_k(\tilde{X})) \geq \rho \right\}
\]

Here, \( \rho \in (0, 1) \) is some empirical thresholding level. We will illustrate this principle on real data analysis in Section 7.2. Under the general deterministic setting, the accurate (or optimal) estimation of tensor rank may be much more challenging and we leave it as future work.

6. Extensions to General Order-d Tensors. While our previous sections mainly focus on order-3 tensor estimation, our results can be generalized to the order-d low-rank tensor estimation with the key ideas outline in this section. First, the constraint set \( C, \) RCG condition and noise quantity \( \xi \) can be defined similarly by replacing the order-3 tensor with the general low-rank order-d tensors; second, the local convergence analysis can be similarly conducted as Theorem 3.1. Define

\[
E^{(t)} := \min_{R_k \in \mathcal{O}_{p_k}, r_k} \left\{ \sum_{k=1}^{d} \left\| U_k^{(t)} - U_k^* R_k \right\|_F^2 + \left\| S^{(t)} - \left[ S^*; R_1^T, \ldots, R_d^T \right] \right\|_F^2 \right\}.
\]

We can build the equivalence between \( E^{(t)} \) and

\[
\left\| \chi^{(t)} - \chi^* \right\|_F^2 + \frac{a}{2} \sum_{k=1}^{d} \left\| U_k^{(t)T} U_k^{(t)} - b^2 I_{r_k} \right\|_F^2
\]

by setting \( b \propto \tilde{\lambda}^{1/(d+1)} \) and \( a \propto \tilde{\lambda} \). Then, we can establish the following theoretical guarantee under a good initialization:

**Theorem 6.1 (Informal)** Suppose \( L \) satisfies RCG(\( \alpha, \beta, C \)) and assume \( \kappa, \alpha, \beta \) are constants. Assume \( \chi^* = [S^*; U_1^*, \ldots, U_d^*] \) and the initialization \( \chi^{(0)} = [S^{(0)}; U_1^{(0)}, \ldots, U_d^{(0)}] \) satisfy

\[
U_k^{*T} U_k = U_k^{(0)T} U_k^{(0)} = b^2 I_{r_k}, \quad U_k^*, U_k^{(0)} \in \mathcal{C}_k, k = 1, \ldots, d; \quad S^*, S^{(k)} \in \mathcal{C}_S.
\]

Suppose the initialization error satisfies \( \left\| \chi^{(0)} - \chi^* \right\| \leq c_d \tilde{\lambda}^2 \) and the signal-noise-ratio satisfies \( \Delta^2 \geq C_d \xi^2 \). Then, by taking step size \( \eta \leq c_d b^{-2d} \), the output of Algorithm 3.1 satisfies

\[
\left\| \chi^{(t)} - \chi^* \right\|_F^2 \leq C_d \left( \xi^2 + (1 - c_d \eta)^t \left\| \chi^{(0)} - \chi^* \right\|_F^2 \right).
\]

Here, \( C_d, c_d, C_d', c_d' \) are constants only depending on \( d \).

It is worth mentioning that Theorem 6.1 applies to low-rank matrix estimation (i.e. \( d = 2 \)): suppose \( X^* = U_1^* S^* U_2^{*T} \) and \( X^{(0)} = U_1^{(0)T} S^{(0)} U_2^{(0)T} \) for \( U_k, U_k^{(0)} \in \mathbb{R}^{p_k \times r}, S^*, S^{(k)} \in \mathbb{R}^{r \times r} \), we have

\[
\left\| X^{(t)} - X^* \right\|_F^2 \leq C \left( \xi^2 + (1 - c \eta)^t \left\| X^{(0)} - X^* \right\|_F^2 \right).
\]
While this framework is more complicated than necessary since one can always decompose a low-rank matrix as the product of two factor matrices $X = U_1 U_2^\top$ without explicitly introducing the “core matrix” $S \in \mathbb{R}^{r \times r}$ (see our previous discussions in Section 3.3).

Based on Theorem 6.1, we can further extend the minimax optimal bounds for the proposed procedure in each application of Section 4, i.e., Theorems 4.1–4.6 to high-order scenarios. We summarize the results to Table 2.

### 7. Numerical Studies.

#### 7.1. Synthetic Data Analysis.

In this section, we investigate the numerical performance of the proposed methods on the problems discussed in Section 4 with simulated data. We assume the true rank $(r_1, r_2, r_3)$ is known to us and the algorithm only involves two tuning parameters: $a$ and $b$. According to Theorem 3.1, a proper choice of $a$ and $b$ primarily depends on the unknown value $\overline{\lambda}$. In practice, we propose to use the initial estimate $\lambda^{(0)}$ as an approximation of $\lambda^*$, use $\overline{\lambda}^{(0)} = \max_k \| \mathcal{M}_k(\lambda^{(0)}) \|$ as a plug-in estimate of $\overline{\lambda}$, then choose $a = \overline{\lambda}^{(0)}$, $b = (\overline{\lambda}^{(0)})^{1/4}$. We consider the following root mean squared error (RMSE) to assess the estimation accuracy in all settings:

\[
(7.1) \quad \text{Loss}(\hat{\lambda}, \lambda^*) = (p_1 p_2 p_3)^{-1/2} ||\hat{\lambda} - \lambda^*||_F.
\]

Average loss over 100 repetitions are reported in following different scenarios.

#### Tensor Regression.

We investigate the numerical performance of the proposed procedure in low-rank tensor regression discussed in Section 4.2. For all simulation settings, we first generate an $r_1$-by-$r_2$-by-$r_3$ core tensor $S$ with i.i.d. standard Gaussian entries and rescale it as $S = S \cdot \lambda / \min_{k=1}^3 \sigma_{\tau_k}(\mathcal{M}_k(S))$. Here, $\lambda$ quantifies the signal level and will be specified later. Then we generate $U_k$ uniformly at random from the Stiefel manifold $O_{p_k, r_k}$ and calculate the true parameter as $\lambda^* = [S; U_1, U_2, U_3]$. The rescaling procedure here ensures that $\min_k \sigma_{\tau_k}(\mathcal{M}_k(\lambda^*)) \geq \lambda$. Now, we draw a random sample based on the regression model (4.2).

We aim to compare the proposed method (Algorithms 1 and 3) with the initialization estimator (Algorithm 3 solely), Tucker-Regression method\(^2\), and MLE. Since the MLE corresponds to the global minimum of the rank-constrained optimization (2.2) and is often computationally intractable, we instead consider a warm-start gradient descent estimator, i.e., performing Algorithm 1 starting from the true parameter $\lambda^*$. We expect that the output of this procedure can well approximate MLE. We implement all four procedures under two settings: (a) $p_1 = p_2 = p_3 = p = 30$, $r_1 = r_2 = r_3 = r = 5$, $\lambda = 2$, $\sigma = 1$, $n$ varies from 300 to 1000; and (b) $p$ varies from 20 to 50, $r = 3$, $\lambda = 2$, $\sigma = 1$, $n = 1.2 p^{3/2} r$. The results

\(^2\)The implementation is based on (Zhou, Li and Zhu, 2013; Zhou, 2017).
are collected in Figure 1. We see from the left panel that for small sample size ($n \leq 600$), the proposed gradient descent method significantly outperforms the Tucker-Regression and initialization estimator while has larger estimation errors than MLE. When the sample size increases ($n \geq 700$), the performance of the proposed gradient descent and Tucker regression algorithms tend to be as good as MLE. Compared to the initialization, gradient descent achieves a great improvement on the estimation accuracy. The right panel of Figure 1 shows that the gradient descent performs as good as the warm-start gradient descent asymptotically and is significantly better than the initialization and Tucker-Regression estimators.

**Poisson Tensor PCA.** Next, we study the numerical performance of the proposed procedure on Poisson tensor PCA. As mentioned earlier in Section 4.3, we found that the projection steps in Algorithm 1 are not essential to the numerical performance, thus we apply Algorithm 1 without the projection steps there in all numerical experiments for Poisson tensor PCA.

For each experiment, we first generate a random core tensor $S \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ with i.i.d. standard normal entries and random orthogonal matrices $U_k$ uniformly on Stiefel manifold $O_{p_k \times r_k}$. Then we calculate $\bar{X} = S \times_1 U_1 \times_2 U_2 \times_3 U_3$ and rescale it as $X^* = \bar{X} \cdot B / \|\bar{X}\|_{\infty}$ to ensure that each entry of $X^*$ is bounded by $B$. Now, we generate a count tensor $Y \in \mathbb{N}^{p_1 \times p_2 \times p_3}$: $Y_{jkl} \sim \text{Poisson}(I \exp(X^*_{jkl}))$ independently, and aim to estimate the low-rank tensor $X^*$ based on $Y$. In addition to the proposed method, we also consider the baseline methods of Poisson-HOSVD and Poisson-HOOI that perform HOSVD and HOOI on log $(((Y + 1/2)/I))$ (i.e., Algorithm 4).

First, we fix $p = 50, r = 5$, vary the intensity value $I$, and study the effect of $I$ to the numerical performance. As we can see from Figure 2, for low intensity, the gradient method is significantly better than two baselines (left panel); for high intensity, three methods are comparable while the Poisson gradient descent is the best (right panel). Next, we study their performance for different tensor dimensions and ranks. In the left panel of Figure 3, we set $r = 5$ and vary $p$ from 30 to 100; in the right panel of Figure 3, we fix $p = 50$ and vary $r$ from 5 to 15. As one can see, our method significantly outperforms the baselines in all settings. All these simulation results illustrate the benefits of applying gradient descent on the Poisson likelihood function.

**Binomial Tensor PCA.** We generate $X^*$ in the same way as the Poisson tensor PCA settings. Suppose we observe $Y \in \mathbb{N}^{p_1 \times p_2 \times p_3}$ generated from

$$Y_{jkl} \sim \text{Binomial} \left( N_{jkl}, s(X^*_{jkl}) \right),$$

independently.
We take all entries with the same population size (i.e., $N_{ijk} = N$) for simplification. We can see from the simulation results in Figure 4 that a larger population size $N$ yields smaller estimation error. In addition, according to Theorem 3.1, the estimation error in theory is of order $O(p^{-1/2}r^{1/2})$, which matches the trend of estimation error curves in Figure 4.
7.2. Real Data Analysis. In this section, we apply the proposed framework to real data applications in 4D-STEM image denoising. An additional real data example on click-through prediction is postponed to Appendix B in the supplementary materials.

The 4D-scanning transmission electron microscopy (4D-STEM) is an important technique in modern material science that has been used to detect local material composition of structures such as films, defects and nanostructures (Krivanek, Dellby and Lupini, 1999; Yankovich et al., 2016). In 4D-STEM imaging technology, a focused probe is usually rastered across part of the specimen and an X-ray and/or electron energy loss spectrum is recorded at each probe position, generating a series of photon-limited images. The data generated from 4D-STEM technique are typically order-4 tensors with approximate periodic structures, as a focused probe is located on a 2-D grid and one 2-D image is generated for each probe position (see Yankovich et al. (2016) for more details). Due to the physical conditions, the observable images are often photon-limited, highly noisy, and in the form of count matrices (see the second row of Figure 5 for an example). A sufficient imaging denoising is often a crucial first step before the subsequent procedures.

We aim to illustrate the merit of the proposed method through denoising of data in 4D-STEM experiments. Specifically, we collect 160 images generated from a row of electron probe positions. Since the resolution of each image is $183 \times 183$, the data images can be stacked into a non-negative tensor of size $160 \times 183 \times 183$. We assume the observational images $Y$ are generated from Poisson distribution $Y_{ijk} \sim \text{Poisson}(\exp(X_{ijk}^*))$. Our goal is to recover the original images based on the photon-limited observation $Y$. Since $Y$ is sparse ($\approx 88\%$ pixels are zero), we take the pre-initializer $\tilde{X} = \log(Y + 1/30)$ and take the input rank according to the (5.3) with $\rho = 0.98$. We apply the proposed gradient descent (Algorithms 1 and 4) with the rank estimation $(\hat{r}_1, \hat{r}_2, \hat{r}_3) = (44, 36, 34)$ to obtain the estimator $\hat{X}$, then calculate $\exp(\hat{X})$ as the collection of denoised images. We also denoise these images one by one via the matrix Procrustes flow (Park et al., 2018), a variant of the matrix-version gradient descent method. The original, observational, and recovered images are provided in Figure 5. In addition, we calculate the recovery loss for each of the 160 images, i.e., $\|\exp(X_{i::}^*) - \exp(\hat{X}_{i::})\|_F / \|\exp(X_{i::}^*)\|_F$, and the averaged recovery loss (and standard error) of matrix and tensor methods are 0.861 (0.183) and 0.303 (0.054), respectively. One can clearly see the advantage of the proposed tensor method that utilizes the tensor structure of the whole set of images.

8. Discussions. In this paper, we introduce a non-convex optimization framework for the generalized tensor estimation. Compared to the convex relaxation methods in the literature, the proposed scheme is computationally efficient and achieves desirable statistical error rate under suitable initialization and signal-to-noise ratio conditions. We apply the proposed framework on several problems, including sub-Gaussian denoising, tensor regression, Poisson and binomial tensor PCA. We can show that the proposed gradient descent procedure achieves the minimax optimal rate of estimation error under these statistical models.

---

3Simultaneously denoising the order-4 image data requires extremely large memory and computation source. Thus we focus on one row of images. It is also common to perform row-wise image denoising in 4D-STEM imaging analysis (Yankovich et al., 2016).

4This algorithm also requires the specification of matrix rank. We empirically choose $\hat{r} = \arg\min\left\{ r : \sum_{i=1}^{pu} \sigma^2_r(X) / \sum_{i=1}^{pu} \sigma^2_r(X) \geq 0.98 \right\}$ for each slice of the pre-initializer $X = X_{i::}$.
In addition to the above-mentioned problems, the proposed framework can incorporate a broader range of settings. For example, the developed result is applicable to solve the noisy tensor completion problem (Xia, Yuan and Zhang, 2017; Montanari and Sun, 2018; Shah and Yu, 2019; Cai et al., 2019), which aims to recover the low-rank tensor $X^*$ based on a number of noisy observable entries, say $\{Y_{ijk} = X^*_{ijk} + Z_{ijk}\}_{(i,j,k) \in \Omega}$, where $\Omega$ is a subset of indices.

Another example is binary tensor PCA (Wang and Li, 2018), where the central goal is to factorize the tensor from 0-1 valued observations. Suppose one observes $Y_{ijk} \sim \text{Bernoulli}(P_{ijk})$ independently, where $P_{ijk} = s(X^*_{ijk})$, $X^*$ is low-rank, and $s(\cdot)$ is some link function. Then the proposed projected gradient descent method can be applied to estimate $X^*$ with provable guarantees.

Community detection in social network has attracted enormous recent attention. Although most of the existing results focused on a single-layer of network, the multilayer network, i.e., the connections between different nodes are reflected in multiple modalities, also commonly appear in practice (Han, Xu and Airoldi, 2015; Lei, Chen and Lynch, 2019; Pensky et al., 2019). Consider a stack of multilayer network data with shared community structure. It is reasonable to assume that the adjacency tensor $A$ has a low-rank tensor structure: $A \sim \text{Bernoulli}(\lambda^*)$ independently, where $\lambda^* = [S^*, Z^*, T^*]$, $Z^*$ is the latent space of
nodes features (or the indicator matrix for the community that each node belongs to), and $T^*$ models the trend along the time. Then the community detection for multilayer networks essentially becomes the generalized tensor estimation problem.

In addition to the standard linear regression model discussed in Section 4.2, the proposed framework can be applied to a range of generalized tensor regression problems. Recall that the classical generalized linear model focuses on an exponential family, where the response $y_i$ satisfies the following density or probability mass function (Nelder and Wedderburn, 1972),

\begin{equation}
\label{eq:glm}
p(y_i|\theta_i, \phi) = \exp \left\{ \frac{y\theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi) \right\}.
\end{equation}

Here, $a, b, c$ are prespecified functions determined by the problem; $\theta_i$ and $\phi > 0$ are natural and dispersion parameters, respectively. For the generalized tensor regression, it is natural to relate the tensor covariate and response (Zhou, Li and Zhu, 2013) via

\begin{equation}
\label{eq:tensor-glm}
\mu_i = \mathbb{E}(y_i|\mathcal{X}^*) \quad g(\mu_i) = \langle A_i, \mathcal{X}^* \rangle,
\end{equation}

where $g(\cdot)$ is a link function. To estimate $\mathcal{X}^*$, we can apply the proposed Algorithm 3 on the negative log-likelihood function

$$
\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi),
$$

where $\theta_i$ is determined by (8.1) and (8.2).

Some other possible applications of the proposed framework include the high-order interaction pursuit (Hao, Zhang and Cheng, 2019), generalized regression among multiple modes (Xu, Hu and Wang, 2019), mixed-data-type tensor data analysis (Baker, Tang and Allen, 2019), etc. In all these problems, by exploring the log-likelihood of data and the domain $C$ that satisfies RCG condition, the proposed projected gradient descent can be applied and the theoretical guarantees can be developed based on the proposed framework.

Acknowledgement. The authors thank Paul Voyles and Chenyu Zhang for providing the 4D-STEM dataset and for helpful discussions. The research of R. H. and A. R. Z. was supported in part by NSF DMS-1811868, NSF CAREER-1944904, and NIH R01-GM131399. The research of R. W. was supported in part by AFOSR FA9550-18-1-0166, DOE DE-AC02-06CH11357, NSF OAC-1934637, and NSF DMS-1930049. The research of R. H. was also supported in part by a RAship from Institute for Mathematics of Data Science at UW-Madison.

REFERENCES

AHMED, A., RECHT, B. and ROMBERG, J. (2013). Blind deconvolution using convex programming. *IEEE Transactions on Information Theory* 60 1711–1732.

ANANDKUMAR, A., HSU, D. and KAKADE, S. M. (2012). A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory* 33–1.

ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research* 15 2773–2832.

ARROYO, J., ATHREYEA, A., CAPE, J., CHEN, G., PRIEBE, C. E. and VOGELSTEIN, J. T. (2019). Inference for multiple heterogeneous networks with a common invariant subspace. *arXiv preprint arXiv:1906.10026*.

BAHADORI, M. T., YU, Q. R. and LIU, Y. (2014). Fast multivariate spatio-temporal analysis via low rank tensor learning. In *Advances in neural information processing systems* 3491–3499.

BAKER, Y., TANG, T. M. and ALLEN, G. I. (2019). Feature Selection for Data Integration with Mixed Multi-view Data. *arXiv preprint arXiv:1903.11232*.

BARAK, B. and MOITRA, A. (2016). Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory* 417–445.
Bi, X., Qu, A. and Shen, X. (2018). Multilayer tensor factorization with applications to recommender systems. *The Annals of Statistics* **46** 3308–3333.

Birgé, L. (2001). An alternative point of view on Lepski’s method. *Lecture Notes-Monograph Series* 113–133.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Cai, T. T., Li, X. and Ma, Z. (2016). Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics* **44** 2221–2251.

Cai, T. T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics* **46** 60–89.

Cai, C., Li, G., Poor, H. V. and Chen, Y. (2019). Nonconvex Low-Rank Tensor Completion from Noisy Data. In *Advances in Neural Information Processing Systems* 1861–1872.

Candes, E. J., Li, X. and Soltanolkotabi, M. (2015). Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory* **61** 1985–2007.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE* **98** 925–936.

Candes, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory* **57** 2342–2359.

Candes, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* **9** 717.

Cao, Y. and Xie, Y. (2015). Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing* **64** 1609–1620.

Cao, Y., Zhang, A. and Li, H. (2019). Multi-sample estimation of bacterial composition matrix in metagenomics data. *Biometrika*.

Chen, W.-K. (2019). Phase transition in the spiked random tensor with rademacher prior. *The Annals of Statistics* **47** 2734–2756.

Chen, Y. and Candes, E. (2015). Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems* 739–747.

Chen, Y. and Chi, Y. (2018). Harnessing structures in big data via guaranteed low-rank matrix estimation. *arXiv preprint arXiv:1802.08397*.

Chen, H., Raskutti, G. and Yuan, M. (2019). Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research* **20** 172–208.

Chi, E. C. and Kolda, T. G. (2012). On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications* **33** 1272–1299.

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

Chi, E. C., Gaines, B. R., Sun, W. W., Zhou, H. and Yang, J. (2018). Provable Convex Co-clustering of Tensors. *arXiv:1803.06518* [stat.ME].

De Lathauwer, L., De Moor, B. and VandeWalle, J. (2000a). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications* **21** 1253–1278.

De Lathauwer, L., De Moor, B. and VandeWalle, J. (2000b). On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications* **21** 1324–1342.

Fan, J., Gong, W. and Zhu, Z. (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*.

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J. and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology* **8** e1002606.

Fazel, M. (2002). Matrix rank minimization with applications.

Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J. R., Domogala, D., Chase, J., Leff, J. W., Vázquez-Baeza, Y., González, A., Knight, R. et al. (2014). Temporal variability is a personalized feature of the human microbiome. *Genome biology* **15** 531.

Friedland, S. and Lim, L.-H. (2018). Nuclear norm of higher-order tensors. *Mathematics of Computation* **87** 1255–1281.

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

Guhaniyogi, R., Qamar, S. and Dunson, D. B. (2017). Bayesian tensor regression. *The Journal of Machine Learning Research* **18** 2733–2763.

Guo, W., Kotsia, I. and Patras, I. (2012). Tensor learning for regression. *IEEE Transactions on Image Processing* **21** 816–827.

Hall, E. C., Raskutti, G. and Willett, R. (2016). Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*. 
HAN, Q., XU, K. and AIROLDI, E. (2015). Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning* 1511–1520.

HAO, B., ZHANG, A. and CHENG, G. (2019). Sparse and Low-rank Tensor Estimation via Cubic Sketchings. *The Annals of Statistics* revision under review.

HENRIQUES, R. and MADEIRA, S. C. (2019). Triclustering algorithms for three-dimensional data analysis: A comprehensive survey. *ACM Computing Surveys (CSUR)* 51 95.

HILLAR, C. J. and LIM, L.-H. (2013). Most tensor problems are NP-hard. *Journal of the ACM (JACM)* 60 45.

HOPF, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics* 9 1169.

HONG, D., KOLDA, T. G. and DUERSCH, J. A. (2018). Generalized canonical polyadic tensor decomposition. *arXiv preprint arXiv:1808.07452*.

HOPKINS, S. B., SHI, J. and STEURER, D. (2015). Tensor principal component analysis via sum-of-square proofs. In *Proceedings of the 28th Conference on Learning Theory, COLT* 3–6.

JAVANMARD, A., MONTANARI, A. et al. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics* 46 2593–2622.

JIANG, X., RASKUTTI, G. and WILLETT, R. (2015). Minimax optimal rates for Poisson inverse problems with physical constraints. *IEEE Transactions on Information Theory* 61 4458–4474.

JOHNDROW, J. E., BHATTACHARYA, A. and DUNSON, D. B. (2017). Tensor decompositions and sparse log-linear models. *The Annals of Statistics* 45 1–38.

JOLLiffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis* 129–155. Springer.

KESHIHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from noisy entries. *Journal of Machine Learning Research* 11 2057–2078.

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM review* 51 455–500.

KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* 39 2302–2329.

KRIVANEK, O., DELLBY, N. and LUPINI, A. (1999). Towards sub-A electron beams. *Ultramicroscopy* 78 1–11.

KROONENBERG, P. M. (2008). *Applied multiway data analysis*. John Wiley & Sons.

LEI, J., CHEN, K. and LYNCH, B. (2019). Consistent community detection in multi-layer network data. *Biometrika*.

LESIEUR, T., MIOLANE, L., LELARGE, M., KRZAKALA, F. and ZDEBOROVÁ, L. (2017). Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)* 511–515. IEEE.

LI, N. and LI, B. (2010). Tensor completion for on-board compression of hyperspectral images. In *2010 IEEE International Conference on Image Processing* 517–520. IEEE.

LI, L. and ZHANG, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association* 1–16.

LI, X., XU, D., ZHOU, H. and LI, L. (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences* 10 520–545.

LIU, J., MUSIALSKI, P., WONKA, P. and YE, J. (2013). Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 208–220.

LUBICH, C., ROWEDETER, T., SCHNEIDER, R. and VANDERHEYDEN, B. (2013). Dynamical approximation by hierarchical Tucker and tensor-train tensors. *SIAM Journal on Matrix Analysis and Applications* 34 470–494.

MA, Z. and MA, Z. (2017). Exploration of Large Networks via Fast and Universal Latent Space Model Fitting. *arXiv preprint arXiv:1705.02372*.

MCMAHAN, H. B., HOLT, G., SCULLY, D., YOUNG, M., EBNER, D., GRADI, J., NIE, L., PHILLIPS, T., DAVYDOD, E., GOLOVIN, D. et al. (2013). Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* 1222–1230. ACM.

MONTANARI, A., REICHMAN, D. and ZEITOUNI, O. (2017). On the Limitation of Spectral Methods: From the Gaussian Hidden Clique Problem to Rank One Perturbations of Gaussian Tensors. *IEEE Transactions on Information Theory* 63 1572–1579.

MONTANARI, A. and SUN, N. (2018). Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics* 71 2381–2425.

NELDER, J. A. and WEDDERBURN, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* 135 370–384.

NESTEROV, Y. (1998). Introductory lectures on convex programming volume i: Basic course. *Lecture notes* 3 5.

OYMAK, S., JALALI, A., FAZEL, M., ELDAR, Y. C. and HASSIBI, B. (2015). Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory* 61 2886–2908.

PAK, D., KRYILLIDIS, A., CARAMANIS, C. and SANGHAVI, S. (2018). Finding low-rank solutions via non-convex matrix factorization, efficiently and provably. *SIAM Journal on Imaging Sciences* 11 2165–2204.
Pensky, M., Zhang, T. et al. (2019). Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics* 13 678–709.

Perry, A., Wein, A. S. and Bandeira, A. S. (2016). Statistical limits of spiked tensor models. *arXiv preprint arXiv:1612.07728*.

Raskutti, G., Yuan, M., Chen, H. et al. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics* 47 1554–1584.

Rauhut, H., Schneider, R. and Stojanac, V. (2015). Tensor completion in hierarchical tensor representations. In *Compressed sensing and its applications* 419–450. Springer.

Rauhut, H., Schneider, R. and Stojanac, Z. (2017). Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications* 523 220–262.

Recht, B., Fazel, M. and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review* 52 471–501.

Richard, E. and Montanari, A. (2014). A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems* 2897–2905.

Salmon, J., Harmany, Z., Deledalle, C.-A. and Willett, R. (2014). Poisson noise reduction with non-local PCA. *Journal of mathematical imaging and vision* 48 279–294.

Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association* 110 1646–1657.

Shah, D. and Yu, C. L. (2019). Iterative Collaborative Filtering for Sparse Noisy Tensor Estimation. *arXiv preprint arXiv:1908.01241*.

Shan, L., Lin, L., Sun, C. and Wang, X. (2016). Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization. *Electronic Commerce Research and Applications* 16 30–42.

Shi, P., Zhou, Y. and Zhang, A. (2018). High-dimensional Log-Error-in-Variable Regression with Applications to Microbial Compositional Data Analysis. *arXiv preprint arXiv:1811.11799*.

Signoretto, M., van de Plas, R., De Moor, B. and Suykens, J. A. (2011). Tensor versus matrix completion: A comparison with application to spectral data. *IEEE Signal Processing Letters* 18 403–406.

Sun, W. W. and Li, L. (2016). Sparse Low-rank Tensor Response Regression. *arXiv preprint arXiv:1609.04523*.

Sun, R. and Luo, Z.-Q. (2015). Guaranteed matrix completion via nonconvex factorization. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on* 270–289. IEEE.

Sun, W. W., Lu, J., Liu, H. and Cheng, G. (2017). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 899–916.

Tibsiriani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Timmerman, K. and Nowak, R. D. (1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Transactions on Information Theory* 45 846–842.

Tomikama, R. and Suzuki, T. (2013). Convex tensor decomposition via structured Schatten norm regularization. In *Advances in neural information processing systems* 1313–1339.

Tomikama, R., Suzuki, T., Hayashi, K. and Kashima, H. (2011). Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems* 972–980.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M. and Recht, B. (2016). Low-rank Solutions of Linear Matrix Equations via Procrustes Flow. In *International Conference on Machine Learning* 964–973.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wang, M., Fischer, J. and Song, Y. S. (2017). Three-way clustering of multi-tissue multi-individual gene expression data using constrained tensor decomposition. *bioRxiv* 229245.

Wang, M. and Li, L. (2018). Learning from Binary Multiway Data: Probabilistic Tensor Decomposition and its Statistical Optimality. *arXiv preprint arXiv:1811.05076*.

Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems* 713–723.

Wen, Z., Yin, W. and Zhang, Y. (2012). Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation* 4 333–361.

Willett, R. M. and Nowak, R. D. (2007). Multiscale Poisson intensity and density estimation. *IEEE Transactions on Information Theory* 53 3171–3187.

Wilmott, J. R. and Shkolnikov, V. (2006). Human mortality database, available at: http://www.mortality.org.

Xia, D. and Yuan, M. (2017). On Polynomial Time Methods for Exact Low Rank Tensor Completion. *arXiv preprint arXiv:1702.06980*.

Xia, D., Yuan, M. and Zhang, C.-H. (2017). Statistically Optimal and Computationally Efficient Low Rank Tensor Completion from Noisy Entries. *arXiv preprint arXiv:1711.04934*. 
Xu, Z., Hu, J. and Wang, M. (2019). Generalized tensor regression with covariates on multiple modes. *arXiv preprint arXiv:1910.09499*.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* 1564–1599.

Yankovich, A. B., Zhang, C., Oh, A., Slater, T. J., Azough, F., Freer, R., Haigh, S. J., Willett, R. and Voyles, P. M. (2016). Non-rigid registration and non-local principle component analysis to improve electron microscopy spectrum images. *Nanotechnology* 27 364001.

Yokota, T., Lee, N. and Cichocki, A. (2016). Robust multilinear tensor rank estimation using higher order singular value decomposition and information criteria. *IEEE Transactions on Signal Processing* 65 1196–1206.

Yonel, B. and Yazici, B. (2020). A Deterministic Convergence Framework for Exact Non-Convex Phase Retrieval. *arXiv preprint arXiv:2001.02855*.

Yu, M., Wang, Z., Gupta, V. and Kolar, M. (2018). Recovery of simultaneous low rank and two-way sparse coefficient matrices, a nonconvex approach. *arXiv preprint arXiv:1802.06967*.

Yuan, M. and Zhang, C.-H. (2014). On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics* 1–38.

Zhang, A. (2019). Cross: Efficient low-rank tensor completion. *The Annals of Statistics* 47 936–964.

Zhang, A., Cai, T. T. and Wu, Y. (2018). Heteroskedastic PCA: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*.

Zhang, A. and Han, R. (2018). Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association* to appear.

Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and Computational Limits. *IEEE Transactions on Information Theory* 64 7311–7338.

Zhang, A., Luo, Y., Raskutti, G. and Yuan, M. (2019). ISLET: Fast and Optimal Low-rank Tensor Regression via Importance Sketching. *arXiv preprint arXiv:1911.03804*.

Zhao, T., Wang, Z. and Liu, H. (2015). A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems* 559–567.

Zhou, H. (2017). Matlab TensorReg Toolbox Version 1.0. Available online at https://huazhou.github.io/TensorReg/.

Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association* 108 540–552.

Zhu, Z., Li, Q., Tang, G. and Wakin, M. B. (2017). The global optimization geometry of nonsymmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256*.