Convex and Nonconvex Geometries of Symmetric Tensor Factorization

Qiuwei Li and Gongguo Tang
Department of Electrical Engineering, Colorado School of Mines, Golden, CO, USA
Email: qiuli@mines.edu, gtang@mines.edu

***Abstract*—Tensors provide natural representations for massive multi-mode datasets and tensor methods also form the backbone of many machine learning, signal processing, and statistical algorithms. This work develops theories and computational methods for guaranteed overcomplete, non-orthogonal symmetric tensor factorization using convex and nonconvex optimizations. In particular, we show when the symmetric tensor factors are uniformly sampled from the unit sphere, they are provably recoverable using convex atomic norm minimization. To design scalable polynomial-time algorithms, we apply low-rank parameterization to reformulate the atomic norm regularized tensor optimization as a nonconvex program. We analyze the optimization landscape of this nonconvex program to ensure (local) convergence of gradient descent algorithms.**

I. Introduction

A d-th order tensor $T\in\bigotimes_{i=1}^d\mathbb{R}^{n_i}$ is an element in the tensor product of Euclidean spaces. Similar as vectors and matrices, a tensor T can be viewed as a multi-way array of numbers

$$T=[T(i_1,i_2,\dots,i_d)]_{i_1\in[n_1],i_2\in[n_2],\dots,i_d\in[n_d]}.$$

Multi-way arrays are useful in representing massive multi-mode datasets, which arise in applications involving multi-view observations, including visual processing [17], collaborative filtering [4], and array signal processing [16]. In addition, the utility of tensors in signal processing and machine learning, especially the symmetric ones, comes from the ability to identify *overcomplete*, *non-orthogonal* factors from tensor data as already suggested by Kruskal's theorem [9]. Unlike matrix decompositions, which are inherently ambiguous due to rotational invariance, that is, for any $U\in\mathbb{R}^{n\times r}$,

$$UU^\top=UQQ^\top U^\top \text{ for any orthogonal matrix } Q,$$

tensor decompositions are essentially unique (up to permutation and sign ambiguities). The process of retrieving tensor factors from tensor data is therefore a well-define inverse problem, known as tensor decomposition/factorization, which is extensively studied across many areas, including machine learning, signal processing, and statistics.

A. Symmetric Tensor Decomposition

In this work, we focus on third order symmetric tensors $T\in(\mathbb{R}^n)^{\otimes 3}$, whose entries $T(i,j,k)$ are invariant with respect to index permutation, *i.e.*, $T(i,j,k)=T(j,k,i)=T(i,k,j)=\cdots$. Any symmetric tensor T can be expressed as a linear combination of unit-norm, rank-one symmetric tensors of the form $\boldsymbol{u}^{\otimes 3}$ with $\boldsymbol{u}\in\mathbb{S}^{n-1}$, where the tensor product is defined via $[\boldsymbol{u}^{\otimes 3}](i,j,k):=\boldsymbol{u}(i)\boldsymbol{u}(j)\boldsymbol{u}(k)$ for $i,j,k\in[n]$. More precisely, any third order symmetric tensor $T^\star\in(\mathbb{R}^n)^{\otimes 3}$ can be expressed in the following form

$$T^\star=\sum_{p=1}^r c_p^\star\boldsymbol{u}_p^{\star\otimes 3} \tag{1}$$

with the total number of terms being r and the factors $\{\boldsymbol{u}_p^\star\}_{p=1}^r$ living on the unit sphere $\mathbb{S}^{n-1}$. Without loss of generality, we also assume that the coefficients $c_p^\star\in\mathbb{R}$ are positive as their signs can always be absorbed into the factors. When the number of rank-1 tensors r is minimal in (1), then r is called the (symmetric) rank of the tensor, and the decomposition (1) becomes a (symmetric) tensor rank decomposition, or Canonical Polyadic (CP) decomposition. Note that a symmetric tensor can also be decomposed into a linear combination of non-symmetric rank-one tensors, which can be used to define the (non-symmetric) tensor rank in a similar manner. Unlike the matrix case, the symmetric tensor rank and the non-symmetric tensor rank might be different for a symmetric tensor.

In this paper, we study symmetric decompositions of (third order) symmetric tensors, which apply naturally to signal processing and machine learning applications. Retrieving such decompositions can be viewed as inverse problems: Given a third order tensor $T^\star$ in the form (1), can we recover the unknown rank-1 factors $\{\boldsymbol{u}_p^\star\}_{p=1}^r$? After retrieving the tensor factors, finding the coefficients $\{c_p^\star\}_{p=1}^r$ reduced to a linear regression problem. The ultimate goals of this work are

- To provide a convex optimization for symmetric tensor decomposition with provable guarantees;
- To develop a scalable polynomial-time algorithm for symmetric tensor decomposition.

B. Our Contributions

Our first result provides convex optimizations for provable symmetric tensor decomposition when the ground truth factors $\{\boldsymbol{u}_p^\star\}_{i=1}^r$ satisfy some mild conditions. In particular, these conditions are satisfied by randomly generated factors. Our approach is based on viewing tensor decomposition as a sparse recovery problem, where the sparsifying dictionary $\{\boldsymbol{u}^{\otimes 3}:\boldsymbol{u}\in\mathbb{S}^{n-1}\}$ is parameterized by points on the unit sphere, and solving an infinite dimensional analog of the ℓ_1 norm minimization. Remarkably, this method is guaranteed to

This work was partially supported by NSF Grants CCF-1464205 and CCF-1704204.

recover an *overcomplete* set of ground truth tensor factors, where the order of rank r can be as large as $n^{17/16} \gg n$. Since this convex formulation is of infinite dimension, which is not directly solvable on computers, our second contribution is applying low-rank parameterization to the atomic norm regularized tensor optimization to develop a polynomial-time algorithms for massive tensor data processing. In spite of the nonconvexity of the resulting factored tensor optimization, in practice, it achieves superior performance even solved by simple local search algorithms, such as gradient descent method. To understand the algorithmic underpinnings of the nonconvex tensor optimization, we analyze the optimization landscapes of a general class of nonconvex tensor optimizations, including tensor decomposition as a special case.

C. Related Works

Tensor Decomposition. Tensor decomposition is one of the multi-mode generalizations of the singular value decomposition for matrices that retrieves the rank-1 tensor factors from tensor data. However, unlike the singular value decomposition for matrices, determining the tensor decomposition for a given tensor is a non-trivial problem that is NP-hard in the worse case [8]. The most widely adopted method for tensor decomposition is the alternating least square method due to its computational efficiency and conceptual simplicity. However, there are much fewer existing global convergence results for the alternating least square method, which might get stuck in local minima or saddle points [5]. The authors of [1] combine the alternating least square approach with a SVD-based initialization scheme, resulting in an algorithm with global convergence guarantees under certain assumptions.

Atomic Norm Minimization. The atomic norm minimization framework provides a principled approach to construct regularizers that encode our prior knowledge about signals [13], [19]. For any signal, its atomic decompositions refer to decompositions in terms of atoms in an atomic set $\mathcal{A}$ that achieve the corresponding atomic norm for that signal. For any atomic norm, it is a fundamental problem to determine conditions under which specific decompositions are atomic decompositions. For example, the singular value decomposition is an atomic decomposition that achieves the matrix atomic norm for the set of unit-norm, rank-one matrices (*i.e.,* the matrix nuclear norm). As shown in [18], for a large class of atomic sets, only decompositions composed of sufficiently *different* atoms are valid atomic decompositions. For tensor decompositions considered in this work, the associate atomic set is the set of unit-norm, rank-one and symmetric tensors. Atomic tensor decompositions are also investigated in [12], [20], but the result of [20] does not apply to overcomplete decompositions, and compared with [12], this work further analyze the optimization landscapes for general tensor recovery to ensure convergence of simple local search algorithms.

Burer-Monteiro Factorization. In the past few years, there have been renewed interest in applying the Burer-Monteiro factorization [3] to low-rank matrix optimization problems for the efficiency and scalability purposes [15]. The underlying idea is that for a low-rank matrix, we can always factorize it as the product of two "tall" matrices, which typically have much fewer variables than the original matrix. In addition, many local search algorithms [10] applied to the Burer-Monteiro reformulations have surprisingly good performance in practice. This phenomena can be understood by analyzing the nonconvex landscape of the factored objective function. Indeed several recent works have shown that the Burer-Monteiro reformulation has no spurious local minima [14], [15], [21], [22]. Such a nice landscape ensures that a range of iterative algorithms such as gradient descent can converge to the global optima with even random initialization [10]. Despite the utility of tensor decomposition in many applications, its widespread adoption in practice has been slow due to its inherent computational intractability. In this work, we apply the Burer-Monteiro factorization idea to tensor decomposition and provide a local convergence result for the simple gradient descent algorithm to solve this problem.

II. The Convex Geometry of Tensor Decomposition

A. Problem Formulation

As mentioned before, this work views tensor decomposition as a sparse recovery problem, where the underlying dictionary $\mathcal{A} = \{\boldsymbol{u}^{\otimes 3} : \boldsymbol{u} \in \mathbb{S}^{n-1}\}$ is (continuously) parametrized by points on the unit sphere. Indeed a tensor $T^\star$ can be expressed as a sparse linear combination of elements from $\mathcal{A}$

$$T^\star = \sum_{p=1}^{r} c_p^\star \boldsymbol{u}_p^{\star\otimes 3},$$

with the number of factors r at most n^3, and in most practical scenarios, r is much smaller than n^3 (but might be much larger than n).

To find the sparse coefficients under the continuously parameterized dictionary $\mathcal{A}$ of infinite number of elements, we view the problem of tensor decomposition as *measure estimation from moments*, where finding the desired sparse coefficients is equivalent to recovering an atomic measure

$$\mu^\star = \sum_{p=1}^{r} c_p^\star \delta(\boldsymbol{u} - \boldsymbol{u}_p^\star) \tag{2}$$

supported on the unit sphere $\mathbb{S}^{n-1}$ from its third order moments

$$T^\star = \int_{\mathbb{S}^{n-1}} \boldsymbol{u}^{\otimes 3} d\mu^\star.$$

This point of view allows us to naturally extend the ℓ_1 minimization in finding sparse representations for finite dictionaries [6] to tensor decomposition, in which the dictionary $\mathcal{A}$ is composed of an infinite number of atoms. More precisely, we recover $\mu^\star$ from the tensor $T^\star$ by solving the following optimization

$$\operatorname*{minimize}_{\mu \in \mathcal{M}(\mathbb{S}^{n-1})} \mu(\mathbb{S}^{n-1}) \text{ subject to } T^\star = \int_{\mathbb{S}^{n-1}} \boldsymbol{u}^{\otimes 3} d\mu \tag{3}$$

where $\mathcal{M}(\mathbb{S}^{n-1})$ is the set of (nonnegative) Borel measures on $\mathbb{S}^{n-1}$. The total measure/mass $\mu(\mathbb{S}^{n-1})$ of the set $\mathbb{S}^{n-1}$ is a generalization of the ℓ_1 norm (or the sum of the entries of a vector) as the measure is assumed to be nonnegative.

B. Connection to The Atomic Norm Approach

For any symmetric tensor T, its atomic norm with respect to $\mathcal{A}$ is defined by:

$$\begin{aligned}\|T\|_{\mathcal{A}} &= \inf\left\{\sum_p \lambda_p : T = \sum_p \lambda_p \boldsymbol{u}_p^{\otimes 3}, \lambda_p > 0, \boldsymbol{u}_p \in \mathbb{S}^{n-1}\right\}\\ &= \inf\{t : T \in t\,\mathrm{conv}(\mathcal{A})\}\end{aligned} \tag{4}$$

where $\mathrm{conv}(\mathcal{A})$ is the convex hull of the atomic set $\mathcal{A}$.

We argue that the two lines in the definition (4) are consistent and are also equivalent to (3) as follows. First, since $\mathrm{conv}(\mathcal{A}) = \{T : T = \int_{\mathbb{K}} \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} d\mu, \mu \in \mathcal{M}(\mathbb{K}), \mu(\mathbb{K}) \leq 1\}$, the optimal value of the second line in the definition (4) is equal to that of (3), and trivially less than or equal to that of the first line in (4). Second, by the famous Carathéodory's convex hull theorem [2], any $n \times n \times n$ symmetric tensor $T \in \mathrm{conv}(\mathcal{A})$ can be expressed as a convex combination of at most $(n+2)(n+1)n/6 + 1$ unit-norm, rank-one symmetric tensors [2], which implies the optimal value of the second line in (3) must be larger than or equal to that of the first line in (3). This argument establishes that the two lines in (4) as well as the measure optimization (3) are equivalent. Therefore, the atomic norm framework and the measure optimization framework are two different formulations of the same problem, with the former setting the stage in the finite dimensional space and the latter in the infinite-dimensional space of measures.

C. Main Result I

Our main results are based on the following assumptions, which are satisfied with high probability by vectors uniformly distributed on the unit sphere $\mathbb{S}^{n-1}$ [1], [7].

Assumption I: Incoherence. The tensor factors are incoherent with a small coherence:

$$\max_{p \neq q} |\langle \boldsymbol{u}_p^\star, \boldsymbol{u}_q^\star \rangle| \leq \frac{\mathrm{polylog}\,(n)}{\sqrt{n}}. \tag{5}$$

Assumption II: Bounded spectral norm. The spectral norm of $U^\star := [\boldsymbol{u}_1^\star \quad \cdots \quad \boldsymbol{u}_r^\star]$ is well-controlled:

$$\|U^\star\| \leq 1 + c\sqrt{\frac{r}{n}} \tag{6}$$

for some numerical constant $c > 0$.

Assumption III: Gram isometry. The Hadamard square of the Gram matrix of $U^\star$ satisfies an isometry condition:

$$\|({U^\star}^\top U^\star) \odot ({U^\star}^\top U^\star) - \mathbf{I}\| \leq \mathrm{polylog}\,(n)\frac{\sqrt{r}}{n}, \tag{7}$$

where $\odot$ is Hadamard product.

Theorem 1. *For a symmetric tensor $T^\star$ in (1) with the rank-one factors $\{\boldsymbol{u}_p^\star\}_{p=1}^r$ satisfying Assumptions I-III and*

$$r = O\left(n^{17/16}/\mathrm{polylog}\,(n)\right), \tag{8}$$

then the true factors $\{\boldsymbol{u}_p^\star\}_{p=1}^r$ can be uniquely recovered by solving (3).

Corollary 1. *For a symmetric tensor $T^\star$ in (1) with the rank-1 factors $\{\boldsymbol{u}_p^\star\}_{p=1}^r$ uniformly sampled from the unit sphere and*

$$r = O\left(n^{17/16}/\mathrm{polylog}\,(n)\right),$$

then the true factors $\{\boldsymbol{u}_p^\star\}_{p=1}^r$ can be uniquely recovered by (3).

D. Proof: Dual Certificate Construction

First, standard Lagrangian analysis shows that the dual problem of (3) is the following semi-infinite program, which has an infinite number of constraints:

$$\begin{aligned}&\underset{Q \in \mathbb{R}^{n\times n\times n}}{\mathrm{maximize}}\ \langle Q, T\rangle\\ &\text{subject to } \langle Q, \boldsymbol{u}^{\otimes 3}\rangle \leq 1, \forall \boldsymbol{u} \in \mathbb{S}^{n-1}.\end{aligned} \tag{9}$$

The polynomial $q(\boldsymbol{u}) := \langle Q, \boldsymbol{u}^{\otimes 3}\rangle = \sum_{i,j,k} Q_{ijk} u_i u_j u_k$ corresponding to a dual feasible solution Q of (9) is called a dual polynomial, which can be used to certify the optimality of a particular decomposition, as demonstrated by the following proposition.

Proposition 1. *Suppose the set of rank-1 tensor factors $\{\boldsymbol{u}_p^{\star\otimes 3}\}_{p=1}^r$ are linearly independent. If there exists a dual feasible solution $Q \in \mathbb{R}^{n\times n\times n}$ of (9) such that the corresponding dual polynomial q satisfies the following* Bounded Interpolation Property (BIP):

$$q(\boldsymbol{u}_p^\star) = 1 \ \text{ for } p = 1, \ldots, r \ \ \text{(Interpolation)}; \tag{10a}$$

$$q(\boldsymbol{u}) < 1 \ \text{ for } \boldsymbol{u} \in \mathbb{S}^{n-1}/S^\star \ \ \text{(Boundedness)}; \tag{10b}$$

where $S^\star = \{\boldsymbol{u}_p^\star\}_{p=1}^r$, then $\mu^\star$ given in (2) is the unique optimal solution to (3).

1) A Pre-certificate: Consider a pre-certificate Q in the form

$$Q = \sum_{p=1}^r \boldsymbol{\alpha}_p^\star \otimes \boldsymbol{u}_p^\star \otimes \boldsymbol{u}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{\beta}_p^\star \otimes \boldsymbol{u}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{u}_p^\star \otimes \boldsymbol{\gamma}_p^\star$$

with the unknown coefficient vectors $\{\boldsymbol{\alpha}_p^\star\}$ to be chosen such that Q satisfies

$$\begin{aligned}Q \times_2 \boldsymbol{u}_p^\star \times_3 \boldsymbol{u}_p^\star &= \boldsymbol{u}_p^\star, \ \text{ for } p = 1, \ldots, r;\\ Q \times_1 \boldsymbol{u}_p^\star \times_3 \boldsymbol{u}_p^\star &= \boldsymbol{u}_p^\star, \ \text{ for } p = 1, \ldots, r;\\ Q \times_1 \boldsymbol{u}_p^\star \times_2 \boldsymbol{u}_p^\star &= \boldsymbol{u}_p^\star, \ \text{ for } p = 1, \ldots, r.\end{aligned} \tag{11}$$

Remark: Clearly, (11) implies the Interpolation property (10a). It remains to show the Boundedness property (10b) for Q to be a valid dual certificate. For this purpose, we check $q(\boldsymbol{u})$ in two regions: the "far" region and the "near" region, which together should cover the entire set $\mathbb{S}^{n-1}$.

2) Far Region: We define the far region as

$$\mathcal{F}(\delta) := \bigcap_{p=1}^r \{\boldsymbol{u} \in \mathbb{S}^{n-1} : |\langle \boldsymbol{u}, \boldsymbol{u}_p^\star \rangle| \leq \delta\} \tag{12}$$

with $\delta \in (0, 1)$ to be chosen later. The far region consists of points that are far away (in the angular sense) from an enlarged support set $\bar{S}^\star = \{\pm \boldsymbol{u}_p^\star : p = 1, \ldots, r\}$.

Lemma 1. *For tensor factors $\{\boldsymbol{u}_p^\star\}_{p=1}^r$ satisfying Assumptions I-III, as long as $r \ll n^{1.25}$, and $r \leq \frac{n}{8\delta c^2}$ for $\delta \leq \frac{1}{8}$, we have the dual polynomial $q(\boldsymbol{u}) < 1$ in $\mathcal{F}(\delta)$.*

3) Near Region: To ensure the union of the far and near regions to cover the entire region $\mathbb{S}^{n-1}$, we define the near region as

$$\mathcal{N}(\delta) := K/\mathcal{F}(\delta) = \bigcup_{p=1}^{r} \{\boldsymbol{u} \in \mathbb{S}^{n-1} : |\langle \boldsymbol{u}_p^\star, \boldsymbol{u} \rangle| \geq \delta\}$$

where the second equality follows from *De Morgan's Law*. One can also treat the whole near region as a union of all individual ones $\mathcal{N}(\delta) = \bigcup_{p=1}^r \mathcal{N}_p(\delta)$ with each individual near region defined as

$$\mathcal{N}_p(\delta) := \{\boldsymbol{u} \in \mathbb{S}^{n-1} : |\langle \boldsymbol{u}_p^\star, \boldsymbol{u} \rangle| \geq \delta\} \tag{13}$$

for $p = 1, \ldots, r$, which consists of all the points that is closed to at least one point in $\tilde{S}^\star$.

Lemma 2. *For tensor factors $\{\boldsymbol{u}_p^\star\}_{p=1}^r$ satisfying Assumptions I-III, if $r \ll n^{1.25}$, we have $q(\boldsymbol{u}) \leq 1$ in the near region $\mathcal{N}(\delta)$ for $\delta = \operatorname{polylog}(n) n^{-0.5 r_c}$, with equality to hold only when $\boldsymbol{u} \in S^\star$.*

Combine Near and Far Regions. The proof is completed by combining Lemma 1 and Lemma 2 and recognizing the union of the far and near regions covers the unit sphere $\mathbb{S}^{n-1}$. To get some intuition, we plot the far and near regions (for $n = 3$ and $r = 2$) in Figure 1.

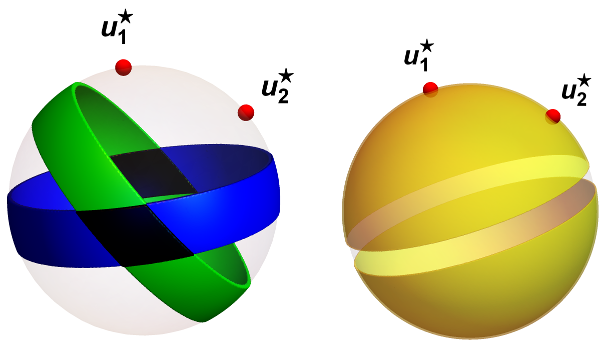

Figure 1. The far region (left) and near region (right): The blue band represents $\{\boldsymbol{u} : |\langle \boldsymbol{u}, \boldsymbol{u}_1^\star \rangle| \leq \delta\}$ that is far away from $\boldsymbol{u}_1^\star$, while the green region $\{\boldsymbol{u} : |\langle \boldsymbol{u}, \boldsymbol{u}_2^\star \rangle| \leq \delta\}$ is the far region associated with $\boldsymbol{u}_2^\star$. The far region is their intersection $\bigcap_{p=1}^2 \{\boldsymbol{u} : |\langle \boldsymbol{u}, \boldsymbol{u}_p^\star \rangle| \leq \delta\}$, consisting of the two black diamonds. The two yellow spherical caps form the near region $\mathcal{N}_1(\delta)$ around the point $\boldsymbol{u}_1^\star$. $\mathcal{N}_2(\delta)$, which is not shown here, consists of another two spherical caps. The union of $\mathcal{N}_1(\delta), \mathcal{N}_2(\delta)$ and the far region $\mathcal{F}(\delta)$ (black diamond) will cover the entire sphere.

III. The Non-convex Geometry of Tensor Decomposition

Theorem 1 for convex geometry shows that under some mild conditions, the identifiability of the rank-1 tensor factors $\{\boldsymbol{u}_p^\star\}_{p=1}^r$ is guaranteed by solving an infinite-dimensional convex optimization (3), which, however, is not directly solvable on a computer. In this section, we apply the Burer-Monteiro low-rank factorization [3], [14], [15], [21], [22] to design fast and scalable algorithm for efficient tensor decomposition.

A. Burer-Monteiro Factorization for Tensors

We can rewrite the symmetric tensor decomposition (1) as

$$T^\star = \sum_{p=1}^{r} c_p^\star \boldsymbol{u}_p^{\star \otimes 3} := \sum_{p=1}^{r} \boldsymbol{x}_p^{\star \otimes 3} := \mathbb{T}(X^\star) \tag{14}$$

with $\boldsymbol{x}_p^\star := c_p^{\star 1/3} \boldsymbol{u}_p^\star$, $X^\star := \begin{bmatrix} \boldsymbol{x}_1^\star & \cdots & \boldsymbol{x}_r^\star \end{bmatrix}$. Clearly, $U^\star$ is a normalized version of $X^\star$.

B. A Reformulation based on Burer-Monteiro Factorization

The proposed method is based on factorizing $T^\star = \mathbb{T}(X)$ for some $X = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_r] \in \mathbb{R}^{n \times \tilde{r}}$ with $\tilde{r} \geq r$.

Proposition 2. *Suppose the tensor decomposition (1) achieves the tensor atomic norm $\|T^\star\|_{\mathcal{A}}$ in (4) and $\tilde{r} \geq r$. Then $\|T^\star\|_{\mathcal{A}}$ is equal to the optimal value of the following optimization:*

$$\operatorname*{minimize}_{X \in \mathbb{R}^{n \times \tilde{r}}} \sum_{p=1}^{\tilde{r}} \|\boldsymbol{x}_p\|_2^3 \ \text{subject to} \ T^\star = \mathbb{T}(X) \tag{15}$$

Proof: For one direction, since $[X^\star \ \mathbf{0}_{n \times (\tilde{r} - r)}]$ is a feasible solution to (15), the optimal value of (15) is no larger than $\sum_{i=1}^r \|\boldsymbol{x}_i^\star\|_2^3$, which is equal to $\|T^\star\|_{\mathcal{A}}$ by the assumption.

For the other direction, suppose an optimal solution of (15) is $X = [\boldsymbol{x}_p]_p \in \mathbb{R}^{n \times \tilde{r}}$. Clearly,

$$T^\star = \sum_{p: \boldsymbol{x}_p \neq 0} \boldsymbol{x}_p^{\otimes 3} = \sum_{p: \boldsymbol{x}_p \neq 0} \|\boldsymbol{x}_p\|_2^3 \boldsymbol{u}_p^{\otimes 3}$$

with $\boldsymbol{u}_p = \boldsymbol{x}_p / \|\boldsymbol{x}_p\|_2$. Then, by (4), we have

$$\|T^\star\|_{\mathcal{A}} \leq \sum_{p: \boldsymbol{x}_p \neq 0} \|\boldsymbol{x}_p\|_2^3 = \sum_{p=1}^{\tilde{r}} \|\boldsymbol{x}_p\|_2^3 = \text{optimum of (15)}.$$

■

Remark: Proposition 2 implies that when an upper bound on r is known, we can solve the nonlinear (and nonconvex) program (15) to compute the tensor atomic norm (and obtain the corresponding decomposition). Further if the ground-truth tensor factors in (1) are uniformly sampled from unit sphere, they can be recovered by the global optimal solution of (15).

Numerical simulations show that solving the nonlinear program (15) using the ADMM method exhibits superior performance. Inspired by the empirical success of applying the Burer-Monteiro method to tensor decomposition [11], we further extend it to general tensor recovery.

C. Extension to General Tensor Recovery

Consider a general symmetric tensor recovery problem

$$T^\star = \arg \operatorname*{minimize}_{T \in S^{n \times n \times n}} f(T) \tag{16}$$

where $S^{n \times n \times n}$ is the space of $n \times n \times n$ symmetric tensors and $f(\cdot)$ is a general (convex in T) objective function. We apply the Burer-Monteiro factorization to (16), solve

$$X^\star \in \arg \operatorname*{minimize}_{X \in \mathbb{R}^{n \times r}} g(X) := f(\mathbb{T}(X)), \tag{17}$$

and set $T^\star = \mathbb{T}(X^\star)$.

Gradient Descent: We propose solving (17) using gradient descent

$$\mathcal{G}(X) = X - \eta \nabla g(X) \tag{18}$$

where η is the step size to be determined later.

D. Main Result II

Notation. Denote $\bar{c} := \max_p c_p^\star$, $\underline{c} := \min_p c_p^\star$, $\frac{\bar{c}}{\underline{c}} = \omega$. Denote the mode-1 flattening of a tensor T as $\mathbb{M}(T)$. For any matrices X, Y of the same size, we define the distance between X and Y by

$$\mathrm{d}(X, Y) = \operatorname*{minimize}_{P \in \Xi} \|X - YP\|_F$$

where Ξ is the set of all permutation matrices of appropriate dimensions. We choose such a distance to accommodate the permutation invariance of tensor decomposition.

Assumption IV: Restricted Well-conditionedness. For every $\mathrm{rank}(T) \le 2r$ and $\mathrm{rank}(D) \le 2r$, we assume

$$m\|D\|_F^2 \le [\nabla^2 f(T)](D, D) \le M\|D\|_F^2 \tag{RIP}$$

Theorem 2. *Under Assumptions I-IV with $r \ll n^{1.25}$, assume the original convex program (16) admits an optimal solution $T^\star = \mathbb{T}(X^\star)$ in (14). We solve its nonconvex formulation (17) via the gradient descent (18) with the initial point X_0 satisfying*

$$\mathrm{d}(X_0, X^\star) \le 0.07 \frac{m}{M} \frac{1}{\omega} \underline{c}^{1/3} \tag{19}$$

and a constant step size $\eta \le \frac{1}{21.6M\|X_0\|^4}$. Let X be the current iterate. Then

$$\mathrm{d}(\mathcal{G}(X), X^\star)^2 \le (1 - 0.26\eta m \underline{c}^{4/3})\, \mathrm{d}(X, X^\star)^2. \tag{20}$$

Remark: Theorem 2 states that if we start from a point with a constant distance (19) from $X^\star$, the gradient descent (18) with a constant sufficiently small step size will converge linearly to the global optimal solution $X^\star$.

E. Proof of Theorem 2

Lemma 3 (Local Regularity Condition). *Under the same settings as in Theorem 2, suppose $P^\star$ is the optimal permutation matrix such that $\mathrm{d}(X, X^\star) = \|X - X^\star P^\star\|_F$. Then*

$$\langle \nabla g(X), X - X^\star P^\star \rangle \ge \frac{1}{2}\eta \|\nabla g(X)\|_F^2 + 0.13 m \underline{c}^{4/3}\, \mathrm{d}(X, X^\star)^2$$

Proof of Theorem 2: The proof is based on Lemma 3.

$$\begin{aligned}
&\mathrm{d}(\mathcal{G}(X), X^\star)^2 \\
&\le \|\mathcal{G}(X) - X^\star P^\star\|_F^2 \\
&= \|\mathcal{G}(X) - X + X - X^\star P^\star\|_F^2 \\
&= \|\mathcal{G}(X) - X\|_F^2 + \|X - X^\star P^\star\|_F^2 + 2\langle \mathcal{G}(X) - X, X - X^\star P^\star \rangle \\
&= \eta^2 \|\nabla g(X)\|_F^2 + \mathrm{d}(X, X^\star)^2 - 2\eta \langle \nabla g(X), X - X^\star P^\star \rangle \\
&\le \eta^2 \|\nabla g(X)\|_F^2 + \mathrm{d}(X, X^\star)^2 \\
&\quad - 2\eta \left(\frac{1}{2}\eta \|\nabla g(X)\|_F^2 + 0.13 m \underline{c}^{4/3}\, \mathrm{d}(X, X^\star)^2 \right) \\
&\le (1 - 0.26 \eta m \underline{c}^{4/3})\, \mathrm{d}(X - X^\star)^2. \qquad \blacksquare
\end{aligned}$$

References

[1] A Anandkumar, Animashree Anandkumar, R Ge, Rong Ge, M Janzamin, and Majid Janzamin. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. *arXiv.org*, 2014.

[2] Alexander Barvinok. *A Course in Convexity*. American Mathematical Soc., 2002.

[3] Samuel Burer and Renato D C Monteiro. A Nonlinear Programming Algorithm for Solving Semidefinite Programs via Low-Rank Factorization. *Mathematical Programming*, 95(2):329–357, February 2003.

[4] Annie Chen. Context-Aware Collaborative Filtering System: Predicting the User's Preference in the Ubiquitous Computing Environment. In *Location- and Context-Awareness*, pages 244–253. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[5] P Comon. Tensor decompositions??state of the art and applications, keynote address in ima conf. *Mathematics in Signal Processing, Warwick, UK*, 2000.

[6] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[7] Olivier Guédon and Mark Rudelson. Lp-moments of random vectors via majorizing measures. *Advances in Mathematics*, 208(2):798–823, 2007.

[8] Christopher J Hillar and Lek-Heng Lim. Most Tensor Problems are NP-Hard. *Journal of the ACM (JACM)*, 60(6):45–39, November 2013.

[9] Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95–138, January 1977.

[10] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.

[11] Qiuwei Li, Ashley Prater, Lixin Shen, and Gongguo Tang. Overcomplete tensor decomposition via convex optimization. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*, pages 53–56. IEEE, 2015.

[12] Qiuwei Li, Ashley Prater, Lixin Shen, and Gongguo Tang. Overcomplete tensor decomposition via convex optimization. *arXiv preprint arXiv:1602.08614*, 2016.

[13] Qiuwei Li and Gongguo Tang. Approximate support recovery of atomic line spectral estimation: A tale of resolution and precision. *arXiv preprint arXiv:1612.01459*, 2016.

[14] Qiuwei Li and Gongguo Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. *arXiv:1611.03060*, 2016.

[15] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. Geometry of factored nuclear norm regularization. *arXiv:1704.01265*, 2017.

[16] Lek-Heng Lim and Pierre Comon. Multiarray signal processing: Tensor decomposition meets compressed sensing. *Comptes Rendus Mecanique*, 338(6):311–320, 2010.

[17] Ji Liu, P Musialski, P Wonka, and Jieping Ye. Tensor Completion for Estimating Missing Values in Visual Data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):208–220, 2013.

[18] Gongguo Tang. Resolution limits for atomic decompositions via Markov-Bernstein type inequalities. In *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pages 548–552, Washington, DC, May 2015. IEEE.

[19] Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

[20] Gongguo Tang and Parikshit Shah. Guaranteed Tensor Decomposition: A Moment Approach. In *International Conference on Machine Learning*, Lille, France, 2015.

[21] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. Global optimality in low-rank matrix optimization. *arXiv preprint arXiv:1702.07945*, 2017.

[22] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of nonsymmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256*, 2017.