Sample Complexity Bounds for Low-Separation-Rank Dictionary Learning

Mohsen Ghassemi, Zahra Shakeri, Waheed U. Bajwa, Anand D. Sarwate
Dept. of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854, USA
Emails: {mohsen.ghassemi, zahra.shakeri, waheed.bajwa, anand.sarwate}@rutgers.edu

***Abstract*—This work addresses the problem of structured dictionary learning for computing sparse representations of tensor-structured data. It introduces a *low-separation-rank dictionary learning* (LSR-DL) model that better captures the structure of tensor data by generalizing the separable dictionary learning model. A dictionary with p columns that is generated from the LSR-DL model is shown to be locally identifiable from noisy observations with recovery error at most ρ given that the number of training samples scales with (# of degrees of freedom in the dictionary) $\times\, p^2\rho^{-2}$.**

I. Introduction

Many data processing tasks such as feature extraction, data compression, classification, signal denoising, image inpainting, and audio source separation make use of data-driven sparse representations of data [1]–[3]. In several applications, these tasks are performed on data samples that are naturally structured as multiway arrays, also known as multidimensional arrays or tensors. Instances of *multidimensional* or *tensor* data include videos, hyperspectral images, tomographic images, and multiple-antenna wireless channels. Sparse representations of these data can allow for significant savings in terms of compression: rather than storing the full signal we can store a sparse vector or coefficients. Traditional approaches to sparse representations of tensor data ignore their multidimensional structure, which results in sparsifying models with a large number of parameters. These larger representation models impact the scalability of applications on computation and/or storage-limited platforms such as smartphones, drones, etc.

Our focus in this paper is on learning of compact models that yield sparse representations of tensor data. To this end, we study *dictionary learning* (DL), an effective and popular data-driven technique for obtaining sparse representations of data [1], [3], for tensor data. The goal in DL is to learn a dictionary $\mathbf{D}$ such that every data sample can be approximated by a linear combination of a few atoms (columns) of $\mathbf{D}$. While DL has been widely studied, traditional DL approaches flatten tensor data and then employ methods designed for vector data [3], [4]. Such simplistic approaches disregard the multidimensional structure in tensor data and result in dictionaries with a large number of parameters. One intuitively expects, however, that dictionaries with smaller number of free parameters that exploit the correlation and structure along different tensor modes are likely to be more efficient with regards to storage requirements, computational complexity, and generalization performance, especially when training data are noisy or scarce.

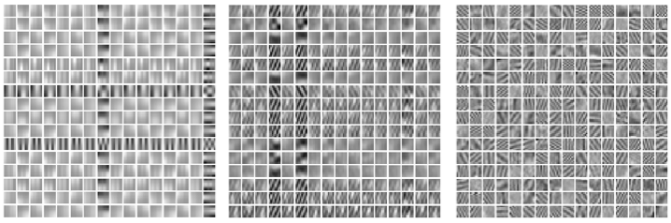

Fig. 1: Dictionary atoms for representing RGB image `Barbara` for separation rank (left-to-right) 1, 4, and 256.

To reduce the number of parameters and better exploit the correlation among different tensor modes, some recent DL works have turned to tensor decompositions such as the Tucker decomposition [5] and CANDECOMP/PARAFAC decomposition (CPD) [6] for learning of structured dictionaries. The idea in *structured DL* for tensor data is to restrict the class of dictionaries to the one imposed by the tensor decomposition under consideration. For example, structured DL based on the Tucker decomposition of N-way tensor data corresponds to the dictionary class in which any dictionary $\mathbf{D} \in \mathbb{R}^{m\times p}$ consists of the Kronecker product [7] of N smaller *subdictionaries* $\{\mathbf{D}_n \in \mathbb{R}^{m_n\times p_n}\}_{n=1}^N$ [8]–[13]. The resulting DL techniques in this instance are interchangeably referred to in the literature as *separable DL* or *Kronecker-structured DL* (KS-DL).

In terms of parameter counting, the advantages of KS-DL for tensor data are straightforward: whereas an unstructured dictionary requires learning and storing $mp = \Pi_{n=1}^N m_n p_n$ parameters, the KS-DL model consists of only $\sum_{n=1}^N m_n p_n$ parameters. Nonetheless, while existing KS-DL methods enjoy lower sample/computational complexity and better storage efficiency over unstructured DL [13], the KS-DL model makes a strong separability assumption among different modes of tensor data, which is often too restrictive [14]. This results in an unfavorable tradeoff between model compactness and representation power. In this paper, we overcome this limitation by proposing and studying a generalization of KS-DL that we interchangeably refer to as *learning a mixture of separable dictionaries* or *low-separation-rank DL* (LSR-DL).

The separation rank of a matrix $\mathbf{A}$ is defined as the minimum number of KS matrices whose sum equals $\mathbf{A}$ [15], [16]. The LSR-DL model interpolates between the under-parameterized separable model (a special case of LSR-DL

model with separation rank 1) and the over-parameterized unstructured model. Figure 1 provides an illustrative example of the usefulness of LSR-DL, in which one learns a dictionary with a small separation rank: while KS-DL learns dictionary atoms that cannot reconstruct diagonal structures perfectly because of the abundance of horizontal/vertical structures within them, LSR-DL also returns dictionary atoms with pronounced diagonal structures as the separation rank increases.

In this work, we study whether the true LSR dictionary underlying tensor data is identifiable by solving a factorization-based LSR-DL problem in which the dictionary is explicitly written as summation of KS matrices and the individual mixture terms have bounded norm. Similar to conventional DL problems, this LSR-DL problem is nonconvex with multiple global minima. Thus, we focus on *local identifiability*, meaning that a local search algorithm initialized close enough to the true dictionary can recover that dictionary. To this end, we show that given a sufficient number of training samples, and under certain assumptions on the generating model and the constraint set, there exists a local minimum of the factorized LSR-DL problem that is within a small neighborhood of the true LSR dictionary generating the data.

A. Our Contributions

This paper studies the information-theoretic limits of the LSR-DL problem, with the hope of informing the future design of computationally efficient algorithms for LSR-DL. Our two main contributions in this regard are as follows:

- We generalize the separable (KS) DL model to a mixture of separable dictionaries, which we call an LSR-DL model. This allows for better representation power than the separable model while maintaining a smaller number of parameters than standard DL.
- We show that in the LSR-DL problem, under certain condition on the problem parameters, $L = \Omega\big(r(\sum_{n=1}^{N} m_n p_n)p^2\rho^{-2}\big)$ samples are sufficient for identifying the true dictionary (underlying Nth-order tensor data of separation rank r) with distortion at most ρ.

B. Related Work

Tensor decompositions [17], [18] have emerged as one of the main sets of tools that help avoid overparameterization of tensor data models in a variety of areas. These include deep learning, collaborative filtering, multilinear subspace learning, source separation, topic modeling, and many other works (see references from recent surveys [19], [20]).

There have been many works that provide theoretical analysis for the sample complexity of the conventional DL problem [21]–[24]. Among these, Gribonval et al. [23] focus on the local identifiability of the true dictionary underlying vectorized data using Frobenius norm as the distance metric. Shakeri et al. [13] extended this analysis for the sample complexity of the KS-DL problem for Nth-order tensor data. That analysis relies on expanding the objective function in terms of subdictionaries and exploiting the coordinate-wise Lipschitz continuity property of the objective function with respect to each subdictionary. While this analysis ensures identifiability of the subdictionaries, it requires the dictionary coefficient vectors to follow the so-called separable sparsity model [12] and does not extend to the LSR-DL problem. In contrast, we provide sample complexity results for the factorization-based LSR-DL problem. Further, our local identifiability result holds for coefficient vectors following either the random sparsity model or the separable sparsity model.

Finally, space constraints prevent us from providing experimental results in this paper; we refer readers to the extended version of this work [25] for a demonstration of the LSR-DL model's usefulness on real data.

II. Background and Problem Statement

A. Notation and Definitions

Underlined bold upper-case ($\underline{\mathbf{A}}$), bold upper-case ($\mathbf{A}$), bold lower-case ($\mathbf{a}$), and regular lower-case letters denote tensors, matrices, vectors, and scalars, respectively. The exception to this convention is the j-th column of a matrix $\mathbf{A}$ which is denoted by $\mathbf{A}_j$. For any integer p, we define $[p] = \{1, 2, \cdots, p\}$. For an $m \times p$ matrix $\mathbf{A}$ and an index set $\mathcal{J} \subseteq [p]$, we denote by $\mathbf{A}_{\mathcal{J}}$ the $|\mathcal{J}| \times p$ matrix containing the columns of $\mathbf{A}$ indexed by $\mathcal{J}$. For vector $\mathbf{v}$, $\|\mathbf{v}\|_0$ and $\|\mathbf{v}\|$ are its ℓ_0 and ℓ_2 norms, while $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_F$, and $\|\mathbf{A}\|_{\text{tr}}$ are the spectral, Frobenius, and trace (nuclear) norms of matrix $\mathbf{A}$, respectively. Further, $\|\mathbf{A}\|_{1\to 2} = \max_j \|\mathbf{A}_j\|_2$ is the max-column norm.

We denote by $(\mathbf{A}_n)_{n=1}^N$ an N-tuple $(\mathbf{A}_1, \cdots, \mathbf{A}_N)$, while $\{\mathbf{A}_n\}_{n=1}^N$ represents the set $\{\mathbf{A}_1, \cdots, \mathbf{A}_N\}$. We often drop the range indicators if they are clear from the context. The Euclidean distance between two tuples of the same size is defined as $\big\|(\mathbf{A}_n)_{n=1}^N - (\mathbf{B}_n)_{n=1}^N\big\|_F \triangleq \sqrt{\sum_{n=1}^N ||\mathbf{A}_n - \mathbf{B}_n||_F^2}$.

We denote the Kronecker product of two matrices by $\otimes$ and the mode-n product between a tensor and a matrix by $\times_n$ [17]. We use $\bigotimes_{n=1}^N \mathbf{A}_n = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_N$ for the Kronecker product of N matrices. For matrix $\mathbf{D}$ with unit-norm columns, its *cumulative coherence* is defined as $\mu_s \triangleq \max_{|\mathcal{J}|\le s} \max_{j\notin\mathcal{J}} \|\mathbf{D}_{\mathcal{J}}^T \mathbf{D}_j\|_1$. We denote by $\mathcal{U}_{m\times p}$ the Euclidean unit sphere: $\mathcal{U}_{m\times p} \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times p} | \|\mathbf{D}\|_F = 1\}$. We also denote the Euclidean sphere with radius α by $\alpha\mathcal{U}_{m\times p}$. The *oblique manifold* in $\mathbb{R}^{m\times p}$ is defined as the manifold of matrices with unit-norm columns: $\mathcal{D}_{m\times p} \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times p} | \forall j \in [p],\ \mathbf{D}_j^T \mathbf{D}_j = 1\}$. The covering number of a set $\mathcal{A}$ with respect to a norm $\|\cdot\|_*$, denoted by $\mathcal{N}_*(\mathcal{A}, \epsilon)$, is the minimum number of balls of $*$-norm radius ϵ needed to cover $\mathcal{A}$.

B. Dictionary Learning Setup

In conventional DL, we assume observations $\mathbf{y} \in \mathbb{R}^m$ are generated according to $\mathbf{y} = \mathbf{D}^0\mathbf{x}^0 + \boldsymbol{\epsilon}$, where $\mathbf{D}^0 \in \mathbb{R}^{m\times p}$ is the underlying dictionary, $\mathbf{x}^0 \in \mathbb{R}^p$ is a randomly generated sparse coefficient vector, and $\boldsymbol{\epsilon} \in \mathbb{R}^m$ is the underlying noise vector. The goal in DL is to recover the underlying dictionary given the noisy observations $\{\mathbf{y}_l\}_{l=1}^L$. One approach is to solve the empirical risk minimization problem

$$\min_{\mathbf{D}\in\mathcal{C}} \ F_{\mathbf{Y}}(\mathbf{D}) = \frac{1}{L}\sum\nolimits_{l=1}^{L} f_{\mathbf{y}_l}(\mathbf{D}) \tag{1}$$

as a proxy for the (unknown) expected risk, where $\mathcal{C} \subseteq \mathbb{R}^{m\times p}$ is the dictionary class, typically selected to be the oblique manifold in unstructured DL, and $f_{\mathbf{y}}(\mathbf{D}) = \inf_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y}-\mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$.

C. Dictionary Learning for Tensor Data

One approach to explicitly account for the tensor structure of data in DL is to use the KS-DL model that is based on Tucker decomposition of tensor data. In the KS-DL model, we assume that observations $\underline{\mathbf{Y}}_l \in \mathbb{R}^{m_1\times\cdots\times m_N}$ are generated according to

$$\underline{\mathbf{Y}}_l = \underline{\mathbf{X}}_l^0 \times_1 \mathbf{D}_1^0 \times_2 \mathbf{D}_2^0 \times_3 \cdots \times_N \mathbf{D}_N^0 + \underline{\mathcal{E}}_l, \qquad (2)$$

where $\mathbf{D}_n^0 \in \mathbb{R}^{m_n\times p_n}$ are generating *subdictionaries*, and $\underline{\mathbf{X}}_l^0$ and $\underline{\mathcal{E}}_l$ are the coefficient and noise tensors, respectively. Equivalently, generating model (2) can be stated for $\mathbf{y}_l \triangleq \mathrm{vec}(\underline{\mathbf{Y}}_l)$ as:

$$\mathbf{y}_l = \left(\mathbf{D}_N^0 \otimes \mathbf{D}_{N-1}^0 \otimes \cdots \otimes \mathbf{D}_1^0\right)\mathbf{x}_l^0 + \boldsymbol{\epsilon}_l, \qquad (3)$$

where $\mathbf{x}_l^0 \triangleq \mathrm{vec}(\underline{\mathbf{X}}_l^0)$ and $\boldsymbol{\epsilon}_l \triangleq \mathrm{vec}(\underline{\mathcal{E}}_l)$. This is the same as the unstructured model $\mathbf{y}_l = \mathbf{D}^0\mathbf{x}^0 + \boldsymbol{\epsilon}_l$ with the additional condition that the generating dictionary is a Kronecker product of N *subdictionaries*. As a result, in the KS-DL problem, the constraint set in (1) becomes $\mathcal{C} = \mathcal{K}_{\mathbf{m},\mathbf{p}}^N$ where $\mathcal{K}_{\mathbf{m},\mathbf{p}}^N \triangleq \{\mathbf{D} \in \mathcal{D}_{m\times p} | \mathbf{D} = \bigotimes_{n=1}^N \mathbf{D}_n,\ \mathbf{D}_n \in \mathbb{R}^{m_n\times p_n}\}$ is the set of KS matrices with unit-norm columns and $\mathbf{m}$ and $\mathbf{p}$ are vectors containing m_n's and p_n's, respectively.

In summary, the structure in tensor data is exploited in the KS-DL model by assuming that the dictionary is separable into subdictionaries for each mode. However, as discussed in the introduction, this separable model is rather restrictive. Instead, we propose a less restrictive model by generalizing the KS-DL model. We begin by defining the *separation rank*.

Definition 1. *The separation rank $\mathfrak{R}_{\mathbf{m},\mathbf{p}}^N(\cdot)$ of a matrix $\mathbf{A} \in \mathbb{R}^{\Pi_n m_n\times\Pi_n p_n}$ is the minimum number r of Nth-order KS matrices $\mathbf{A}^k = \bigotimes_{n=1}^N \mathbf{A}_n^k$ such that $\mathbf{A} = \sum_{k=1}^r \mathbf{A}^k$, where $\mathbf{A}_n^k \in \mathbb{R}^{m_n\times p_n}$.*

Hence, the KS-DL model corresponds to separation rank 1. In contrast, the LSR-DL model is the one in which the separation rank of the underlying dictionary is relatively small so that $1 \leq \mathfrak{R}_{\mathbf{m},\mathbf{p}}(\mathbf{D}^0) \ll \min\{m,p\}$. This generalizes the KS-DL model to a generating dictionary of the form $\mathbf{D}^0 = \sum_{k=1}^r [\mathbf{D}_N^k]^0 \otimes [\mathbf{D}_{N-1}^k]^0 \otimes \cdots \otimes [\mathbf{D}_1^k]^0$, where r is the separation rank of $\mathbf{D}^0$. Consequently, given $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r} \triangleq \{\mathbf{D} \in \mathcal{D}_{m\times p} | \mathfrak{R}_{\mathbf{m},\mathbf{p}}^N(\mathbf{D}) \leq r\}$, the empirical *rank-constrained LSR-DL problem* is

$$\min_{\mathbf{D}\in\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}} F_{\mathbf{Y}}(\mathbf{D}). \qquad (4)$$

To prove identifiability of the true LSR dictionary, the analytical tools at our disposal require the constraint set in (4) to be closed. However, the set $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ is not closed when $N > 2$ and $r > 1$. In that case, we instead analyze (4) with a certain closed subset of $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ (see the discussion in Section III). In our study of LSR-DL (which includes KS-DL as a special case), we use the following correspondence between KS matrices and rank-1 tensors, proved in our earlier work [26], which allows us to leverage techniques and results in the tensor recovery literature to analyze the LSR-DL recovery problem and develop tractable LSR-DL algorithms.

Lemma 1. *Any Nth-order KS matrix $\mathbf{A} = \mathbf{A}_1\otimes\mathbf{A}_2\otimes\cdots\otimes\mathbf{A}_N$ can be rearranged as a rank-1 tensor $\underline{\mathbf{A}}^\pi = \mathbf{a}_N\circ\cdots\circ\mathbf{a}_2\circ\mathbf{a}_1$.*

It follows immediately from Lemma 1 that if $\mathbf{D} = \sum_{k=1}^r \mathbf{D}_1^k \otimes \mathbf{D}_2^k \otimes \cdots \otimes \mathbf{D}_N^k$, then we can rearrange matrix $\mathbf{D}$ into the tensor $\underline{\mathbf{D}}^\pi = \sum_{k=1}^r \mathbf{d}_N^k \circ \mathbf{d}_{N-1}^k \circ \cdots \circ \mathbf{d}_1^k$, where $\mathbf{d}_n = \mathrm{vec}(\mathbf{D}_n)$. Therefore, we have the equivalence $\mathfrak{R}_{\mathbf{m},\mathbf{p}}^N(\mathbf{D}) \leq r \iff \mathrm{rank}(\underline{\mathbf{D}}^\pi) \leq r$. This correspondence highlights a challenge with the LSR-DL problem: finding the rank of a tensor is NP-hard [27], [28] and thus so is finding the separation rank of a matrix. This makes the rank-constrained Problem (4) in its current form (and its variant for $N > 2$, $r > 1$) intractable. To overcome this issue, we introduce a tractable relaxation to the rank-constrained Problem (4) that does not require explicit computation of the tensor rank. The relaxation that we propose here is the *factorization-based LSR-DL* model, in which the LSR dictionary is explicitly written in terms of its subdictionaries. The resulting empirical risk minimization problem is

$$\min_{\{\mathbf{D}_n^k\}:\ \sum_{k=1}^r \bigotimes_{n=1}^N \mathbf{D}_n^k \in \mathcal{D}_{m\times p}} F_{\mathbf{Y}}^{\mathrm{fac}}\left(\{\mathbf{D}_n^k\}\right), \qquad (5)$$

where $F_{\mathbf{Y}}^{\mathrm{fac}}(\{\mathbf{D}_n^k\}) \triangleq \frac{1}{L}\sum_{l=1}^L f_{\mathbf{y}_l}^{\mathrm{fac}}(\{\mathbf{D}_n^k\})$ with $f_{\mathbf{y}}^{\mathrm{fac}}(\{\mathbf{D}_n^k\}) \triangleq \inf_{\mathbf{x}\in\mathbb{R}^p} \left\|\mathbf{y} - \left(\sum_{k=1}^r \bigotimes_{n=1}^N \mathbf{D}_n^k\right)\mathbf{x}\right\|^2 + \lambda\|\mathbf{x}\|_1$ and the terms $\bigotimes_{n=1}^N \mathbf{D}_n^k$ are constrained as $\|\bigotimes_{n=1}^N \mathbf{D}_n^k\|_F \leq c$ for some positive constant c when $N > 2$ and $r > 1$.

D. Generating Model

Here, we describe the generating model that we consider in this work. Let $\mathbf{D}^0 \in \mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ be the underlying dictionary. Each training sample $\underline{\mathbf{Y}} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_N}$ is *independently* generated using a linear combination of $s \ll p$ atoms of dictionary $\mathbf{D}^0$ with added noise: $\mathbf{y} \triangleq \mathrm{vec}(\underline{\mathbf{Y}}) = \mathbf{D}^0\mathbf{x}^0 + \boldsymbol{\epsilon}$ where $\left\|\mathbf{x}^0\right\|_0 \leq s$. Specifically, s atoms of $\mathbf{D}^0$ are selected uniformly at random, defining the support $\mathcal{J} \subset [p]$. Then, we draw a random sparse coefficient vector $\mathbf{x}^0 \in \mathbb{R}^p$ supported on $\mathcal{J}$. We state further assumptions on the distribution of $\mathbf{x}^0$ and $\boldsymbol{\epsilon}$ that are similar to the ones in [23] and [13].

Assumption 1 (Coefficient Distribution)**.** *Consider a random variable $x \in \mathbb{R}$. Define $\mathbf{s}^0 = \mathrm{sgn}(\mathbf{x}^0)$. We assume: **i)** $\mathbb{E}\{\mathbf{x}_{\mathcal{J}}^0[\mathbf{x}_{\mathcal{J}}^0]^T|\mathcal{J}\} = \mathbb{E}\{x^2\}\cdot\mathbf{I}_s$, **ii)** $\mathbb{E}\{\mathbf{s}_{\mathcal{J}}^0[\mathbf{s}_{\mathcal{J}}^0]^T|\mathcal{J}\} = \mathbf{I}_s$, **iii)** $\mathbb{E}\{\mathbf{s}_{\mathcal{J}}^0[\mathbf{x}_{\mathcal{J}}^0]^T|\mathcal{J}\} = \mathbb{E}\{|x|\}\cdot\mathbf{I}_s$, **iv)** $\left\|\mathbf{x}^0\right\|_2 \leq M_x$ with probability 1, **v)** $\min_{j\in\mathcal{J}} |\mathbf{x}_j^0| \geq \underline{x}$ with probability 1.*

Assumption 2 (Noise Distribution)**.** *Consider a random variable $\epsilon \in \mathbb{R}$. **i)** $\mathbb{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\mathcal{J}\} = \mathbb{E}\{\epsilon^2\}\cdot\mathbf{I}_m$, **ii)** $\mathbb{E}\{\mathbf{x}^0\boldsymbol{\epsilon}^T|\mathcal{J}\} = \mathbb{E}\{\mathbf{s}^0\boldsymbol{\epsilon}^T|\mathcal{J}\} = \mathbf{0}$, **iii)** $\|\boldsymbol{\epsilon}\|_2 \leq M_\epsilon$ with probability 1.*

Assumptions 1-iv and 2-iii imply the magnitude of $\mathbf{y}$ is bounded: $\|\mathbf{y}\|_2 \leq M_y$. Next, we define parameters $C_{\min} \triangleq 24\frac{\mathbb{E}\{|x|\}^2}{\mathbb{E}\{x^2\}}\left(\left\|\mathbf{D}^0\right\|_2+1\right)^2\frac{s}{p}\left\|[\mathbf{D}^0]^T\mathbf{D}^0-\mathbf{I}\right\|_F$, $C_{\max} \triangleq \frac{2\mathbb{E}\{|x|\}}{7M_x}\left(1-2\mu_s(\mathbf{D}^0)\right)$, and $\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\{|x|\}}$ for ease of notation. We also make the following standard assumptions:

Assumption 3. *Assume $C_{\min} \leq C_{\max}$, $\lambda \leq \underline{x}/4$, $s \leq \frac{p}{16\left(\|\mathbf{D}^0\|_2+1\right)^2}$, $\mu_s(\mathbf{D}^0) \leq 1/4$, and the noise is relatively small in the sense that $\frac{M_\epsilon}{M_x} < \frac{7}{2}\left(C_{\max}-C_{\min}\right)\bar{\lambda}$.*

In the rest of this paper, we study the problem of recovering the true underlying LSR-DL dictionary by solving the factorization-based LSR-DL Problem (5).

III. Identifiability Result

In this section, we derive the sample complexity required to guarantee with high probability that the true dictionary $\mathbf{D}^0 \in \mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ is identifiable as a solution to the minimization problem in (5). More specifically, we find the number of samples required to guarantee that, for the generating model described in the previous section, there is at least one local minimum $\{[\mathbf{D}^k_n]^*\}$ of the factorization-based LSR-DL Problem (5) such that $\sum\bigotimes[\mathbf{D}^k_n]^*$ is close to the underlying dictionary $\mathbf{D}^0$. This implies that given enough samples, any DL algorithm that converges to a local minimum of this problem can recover $\mathbf{D}^0$ up to a small error if it is initialized close enough to $\mathbf{D}^0$.

Our analysis of the sample complexity of Problem (5) is based on connecting the local minima of Problem (5) to those of Problem (4) and showing that local identifiability guarantees for Problem (4) translate to those for Problem (5). Hence, we first derive sufficient conditions on number of samples required for local identifiability of Problem (4).

Theorem 1. *Consider the DL Problem* (1) *with a compact constraint set $\mathcal{C} \subseteq \mathcal{D}_{m\times p}$ and fix any $u > 0$. Suppose that $\mathbf{D}^0 \in \mathcal{C}$ and that Assumptions 1–3 are satisfied. Assume $\bar{\lambda}C_{\min} < \rho < \bar{\lambda}C_{\max}$ and $\frac{M_\epsilon}{M_x} < \frac{7}{2}(\bar{\lambda}C_{max}-\rho)$. Define positive constants c_0 and ν such that $\mathcal{N}_{1\to2}(\mathcal{C},\varepsilon) = \left(\frac{c_0}{\varepsilon}\right)^\nu$. Suppose the number of samples L satisfies*

$$\frac{L}{\log L} \geq Cp^2\left(c_0\nu+u\right)\frac{M_y^4}{\left(\rho\left(\rho-\bar{\lambda}C_{\min}\right)\mathbb{E}\{x^2\}\right)^2}, \qquad (6)$$

where C is a positive constant, then, with probability no less than $1-e^{-u}$, the objective function $\mathbf{D}\in\mathcal{C}\mapsto F_{\mathbf{Y}}(\mathbf{D})$ has a local minimizer $\mathbf{D}^*$ such that $\left\|\mathbf{D}^*-\mathbf{D}^0\right\|_F \leq \rho$.

The result in Theorem 1 holds for *compact* constraint sets. To apply it to the LSR-DL problem, we need to study the compactness of $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$. Since $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ is a bounded set, according to the Heine-Borel Theorem [29], it is a compact subset of $\mathbb{R}^{m\times p}$ if and only if it is closed. This set can be written as the intersection of the set $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}} = \{\mathbf{D}\in\mathbb{R}^{m\times p}|\mathfrak{R}^N_{\mathbf{m},\mathbf{p}}(\mathbf{D}) \leq r\}$ and the oblique manifold $\mathcal{D}$. To show $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}} = \mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}} \cap \mathcal{D}$ is closed, it suffices to show that $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ and $\mathcal{D}$ are closed. However, we show in the extended version of this work [25] that although the set of KS matrices $\mathcal{L}^{N,1}_{\mathbf{m},\mathbf{p}}$ and the set $\mathcal{L}^{2,r}_{\mathbf{m},\mathbf{p}}$ are closed, the set $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ is not closed in general for $N \geq 3$ and $r \geq 2$. To work around this issue, we consider ${}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}} \triangleq \left\{\mathbf{D}\in\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}|\forall k\in[r],\ \left\|\bigotimes_{n=1}^N\mathbf{D}^k_n\right\|_F \leq c\right\}$, a closed subset of $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$. Note that $\mathcal{K}^N_{\mathbf{m},\mathbf{p}} = {}^c\mathcal{K}^{N,1}_{\mathbf{m},\mathbf{p}}$ and $\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}} = {}^c\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}$.

The sample complexity result in Theorem 1 depends on the covering number of the set of interest. The following lemma bounds the covering number of ${}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$.

Lemma 2. *For the covering number of the set ${}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ with respect to the max-column norm $\|\cdot\|_{1\to2}$, we have the bound $\mathcal{N}_{1\to2}({}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}},\varepsilon) \leq (3rc/\varepsilon)^{r\sum_{n=1}^N m_np_n}$.*

The proofs for Theorem 1 and Lemma 2 can be found in the extended version of this work [25]. Before we present the local identifiability result of factorization-based LSR-DL, we need the following lemma that establishes a bound on the distance of LSR matrices when their factor matrices are ϵ-close.

Lemma 3. *For any two tuples $(\mathbf{A}^k_n)$ and $(\mathbf{B}^k_n)$ such that $\mathbf{A}^k_n, \mathbf{B}^k_n \in \alpha\mathcal{U}_{m_n\times p_n}$ for all $n\in[N]$ and $k\in[r]$, if the distance $\left\|(\mathbf{A}^k_n)-(\mathbf{B}^k_n)\right\|_F \leq \xi$, then $\left\|\sum_{k=1}^r\bigotimes\mathbf{A}^k_n - \sum_{k=1}^r\bigotimes\mathbf{B}^k_n\right\|_F \leq \alpha^N\sqrt{Nr}\xi$.*

The proof of Lemma 3 is also provided in [25].

Theorem 2 (Main Result)**.** *Consider the factorization-based LSR-DL Problem* (5)*. Suppose that the assumptions for Theorem 1 are satisfied. Let the number of samples satisfy sample complexity requirement* (6) *with $\nu = r\sum_{n=1}^N m_np_n$. Then with probability no less than $1-e^{-u}$, the empirical risk objective function $F^{\text{fac}}_{\mathbf{Y}}\left(\{\mathbf{D}^k_n\}\right)$ has a local minimum achieved at $\{[\mathbf{D}^k_n]^*\}$ such that $\left\|\sum\bigotimes[\mathbf{D}^k_n]^* - \mathbf{D}^0\right\|_F \leq \rho$.*

Proof. Let us first establish a connection between the local minima of (5) and those of (4). It is easy to show that any $\mathbf{D}\in{}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ can be written as $\sum_{k=1}^r\bigotimes\mathbf{D}^k_n$ for all $k\in[r]$ and $n\in[N]$ such that, without loss of generality, $\mathbf{D}^k_n \in \alpha\mathcal{U}_{m\times p}$ where $\alpha = \sqrt[N]{c}$. Define

$$\mathcal{C}^{\text{fac}} \triangleq \left\{(\mathbf{D}^k_n)\middle|\sum\bigotimes\mathbf{D}^k_n \in {}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}, \forall k,n : \mathbf{D}^k_n\in\alpha\mathcal{U}_{m\times p}\right\}.$$

Since $\mathbf{D}^*\in{}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$, there is a $([\mathbf{D}^k_n]^*)\in\mathcal{C}^{\text{fac}}$ such that $\mathbf{D}^* = \sum\bigotimes[\mathbf{D}^k_n]^*$. According to Lemma 3, for any $\{\mathbf{D}^k_n\}\in\mathcal{C}^{\text{fac}}$ and any $\xi'>0$, if $\left\|(\mathbf{D}^k_n)-([\mathbf{D}^k_n]^*)\right\|_F \leq \xi'$, it follows that $\left\|\sum\bigotimes\mathbf{D}^k_n - \sum\bigotimes[\mathbf{D}^k_n]^*\right\|_F \leq \alpha^N\sqrt{Nr}\xi' = c\sqrt{Nr}\xi'$. Since $\mathbf{D}^*$ is a local minimizer of (4), there exists a positive ξ such that for all $\mathbf{D}\in{}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ satisfying $\|\mathbf{D}-\mathbf{D}^*\|_F\leq\xi$, we have $F_{\mathbf{Y}}(\mathbf{D}^*) \leq F_{\mathbf{Y}}(\mathbf{D})$. If we choose ξ' small enough such that $c\sqrt{Nr}\xi' \leq \xi$, then for any $(\mathbf{D}^k_n)\in\mathcal{C}^{\text{fac}}$ such that $\left\|(\mathbf{D}^k_n)-([\mathbf{D}^k_n]^*)\right\|_F\leq\xi'$, we have $\left\|\sum\bigotimes\mathbf{D}^k_n-\mathbf{D}^*\right\|_F \leq \xi$ and this means that $F^{\text{fac}}_{\mathbf{Y}}\left(\{\mathbf{D}^k_n\}\right) - F^{\text{fac}}_{\mathbf{Y}}\left(\{[\mathbf{D}^k_n]^*\}\right) = F_{\mathbf{Y}}(\sum\bigotimes\mathbf{D}^k_n) - F_{\mathbf{Y}}(\mathbf{D}^*) \geq 0$. Therefore, $([\mathbf{D}^k_n]^*)$ is a local minimizer of Problem (5).

We established that if there exists a local minimum $\mathbf{D}^*$ of (4) close to $\mathbf{D}^0$, then there is a local minimum $\{[\mathbf{D}^k_n]^*\}$ of (5) such that $\sum\bigotimes[\mathbf{D}^k_n]^*$ is close to $\mathbf{D}^0$. Thus, local identifiability guarantees for Problem (4) directly translate to local identifiability guarantees for Problem (5). That is, if the sample complexity requirement (6) is met, a local minimum $\{[\mathbf{D}^k_n]^*\}$ of (5) is such that $\sum\bigotimes[\mathbf{D}^k_n]^*$ is close to $\mathbf{D}^0$. □

TABLE I: Comparison of known upper bounds on the sample complexity of KS-DL, LSR-DL, and standard DL.

	Best Known	This Paper
KS (Separable Sparsity)	$\Omega(m_n p_n^3 \rho_n^{-2})$ [13]	$\Omega\big((\sum_n m_n p_n) p^2 \rho^{-2}\big)$
KS (Random Sparsity)	$\Omega\big((\sum_n m_n p_n) p^2 \rho^{-2}\big)$ [30]	$\Omega\big((\sum_n m_n p_n) p^2 \rho^{-2}\big)$
LSR	–	$\Omega\big(r(\sum_n m_n p_n) p^2 \rho^{-2}\big)$
Standard	$\Omega(mp^3\rho^{-2})$ [23]	$\Omega(mp^3\rho^{-2})$

IV. Discussion

Here, we have provided a sample complexity upper-bound for local identifiability in the factorization-based LSR-DL problem. We compare our result with the known sample complexity bounds for local identifiability of KS and standard DL problems in Table I. Note that when the separation rank is 1, our result gives a bound on the sample complexity of the KS-DL model as a special case. Unlike previous analysis for the KS-DL model [13], which has sample complexity of $\max_{n\in\{1,\dots,N\}} \Omega(m_n p_n^3 \rho_n^{-2})$, our analysis of the factorized model does not focus on identifiability of the true subdictionaries. However, we do away with the requirement that the dictionary coefficient vectors follow the separable sparsity model: our result does not require any constraints on the sparsity pattern of the coefficient vector.

V. Conclusion

We studied the low-separation-rank model (LSR-DL) to learn structured dictionaries for tensor data. This model bridges the gap between unstructured and separable dictionary learning (DL) models. We show that given $\Omega\big(r(\sum_n m_n p_n) p^2 \rho^{-2}\big)$ data samples, the true dictionary can be locally recovered up to distance ρ. This is a reduction compared to the $\Omega(mp^3\rho^{-2})$ sample complexity of standard DL in [23]. However, a minimax lower bound scaling of $\Omega(p\sum_n m_n p_n \rho^{-2})$ in [12] for KS-DL ($r = 1$) has an $O(p)$ gap with our upper bound. This shows an interesting possible future direction to tighten these bounds.

Acknowledgement

This work is supported in part by the National Science Foundation under awards CCF-1453432 and CCF-1453073, and by the Army Research Office under award W911NF-17-1-0546.

References

[1] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, 2003.

[2] M. Elad, J.-L. Starck, P. Querre, and D. L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *Appl. and Computational Harmonic Anal.*, vol. 19, no. 3, pp. 340–358, 2005.

[3] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, November 2006.

[5] L. R. Tucker, "Implications of factor analysis of three-way matrices for measurement of change," *Problems in Measuring Change*, pp. 122–137, 1963.

[4] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[6] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[7] C. F. Van Loan, "The ubiquitous Kronecker product," *J. Computational and Appl. Math.*, vol. 123, no. 1, pp. 85–100, 2000.

[8] S. Hawe, M. Seibert, and M. Kleinsteuber, "Separable dictionary learning," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2013, pp. 438–445.

[9] F. Roemer, G. Del Galdo, and M. Haardt, "Tensor-based algorithms for learning multidimensional separable dictionaries," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2014, pp. 3963–3967.

[10] C. F. Dantas, M. N. da Costa, and R. da Rocha Lopes, "Learning dictionaries as a sum of Kronecker products," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 559–563, March 2017.

[11] S. Zubair and W. Wang, "Tensor dictionary learning with sparse Tucker decomposition," in *Proc. IEEE 18th Int. Conf. Digital Signal Process. (DSP)*, 2013, pp. 1–6.

[12] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds on dictionary learning for tensor data," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2706–2726, April 2018.

[13] Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "Identifiability of Kronecker-structured dictionaries for tensor data," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 5, pp. 1047 – 1062, 2018.

[14] N. Cressie and H.-C. Huang, "Classes of nonseparable, spatio-temporal stationary covariance functions," *J. American Statistical Association*, vol. 94, no. 448, pp. 1330–1339, 1999.

[15] G. Beylkin and M. J. Mohlenkamp, "Numerical operator calculus in higher dimensions," *Proceedings of the National Academy of Sciences*, vol. 99, no. 16, pp. 10 246–10 251, 2002.

[16] T. Tsiligkaridis and A. O. Hero, "Covariance estimation in high dimensions via Kronecker product expansions," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5347–5360, 2013.

[17] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, August 2009.

[18] I. V. Oseledets, "Tensor-train decomposition," *SIAM J. Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011.

[19] A. Novikov, D. Podoprikhin, A. Osokin, and D. P. Vetrov, "Tensorizing neural networks," in *Proc. Advances in Neural Inform. Process. Syst.*, 2015, pp. 442–450.

[20] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, 2017.

[21] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Proc. 25th Annu. Conf. Learning Theory*, ser. JMLR: Workshop and Conf. Proc., vol. 35, 2014, pp. 1–28.

[22] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries," in *Proc. 27th Annu. Conf. Learning Theory*, ser. JMLR: Workshop and Conf. Proc., vol. 35, no. 1, 2014, pp. 1–15.

[23] R. Gribonval, R. Jenatton, and F. Bach, "Sparse and spurious: Dictionary learning with noise and outliers," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6298–6319, 2015.

[24] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *Appl. and Computational Harmonic Anal.*, vol. 37, no. 3, pp. 464–491, 2014.

[25] M. Ghassemi, Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "Learning mixtures of separable dictionaries for tensor data: Analysis and algorithms," *arXiv preprint arXiv:1903.09284*, 2019.

[26] ——, "STARK: Structured dictionary learning through rank-one tensor recovery," in *Proc. IEEE 7th Int. Workshop Computational Advances in Multi-Sensor Adaptive Processing*, 2017, pp. 1–5.

[27] J. Håstad, "Tensor rank is NP-complete," *J. Algorithms*, vol. 11, no. 4, pp. 644–654, 1990.

[28] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *J. ACM*, vol. 60, no. 6, p. 45, 2013.

[29] W. Rudin, *Principles of mathematical analysis*. McGraw-Hill New York, 1976, vol. 3, no. 4.2.

[30] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, "Sample complexity of dictionary learning and other matrix factorizations," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3469–3486, 2015.