Identification of Kronecker-structured Dictionaries: An Asymptotic Analysis

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

***Abstract*—The focus of this work is on derivation of conditions for asymptotic recovery of Kronecker-structured dictionaries underlying second-order tensor data. Given second-order tensor observations (equivalently, matrix-valued data samples) that are generated using a Kronecker-structured dictionary and sparse coefficient tensors, conditions on the dictionary and coefficient distribution are derived that enable asymptotic recovery of the individual coordinate dictionaries comprising the Kronecker dictionary within a local neighborhood of the true model. These conditions constitute the first step towards understanding the sample complexity of Kronecker-structured dictionary learning for second- and higher-order tensor data.**

I. Introduction

The age of big data has given rise to massive multi-dimensional datasets that need to be processed, stored, and communicated efficiently. In many applications, samples in these datasets can be represented using a *tensor* structure. Developing representation learning techniques that exploit this structural correlation across dimensions will result in more efficient data storage and processing. Our focus in this paper is on two-dimensional data, such as images, that intrinsically possess high spatial correlation. While there exist several works that exploit this correlation for applications such as image recognition and classification [1], [2], relatively less attention has been paid to exploitation of this correlation for representation learning. In particular, dictionary learning (DL) is a (data-driven) representation learning technique that results in sparse representations of data that can then be leveraged for denoising, compression, classification, etc. But traditional DL approaches do not account for the structure of tensor data and instead rely on vectorization for tensor data representation [3]–[5]. In contrast, notice that a second-order tensor data point $\mathbf{Y} \in \mathbb{R}^{m_1 \times m_2}$ can usually be decomposed as $\mathbf{Y} = \mathbf{A}\mathbf{X}\mathbf{B}^\top$, where $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ denotes the sparse coefficient tensor and $\mathbf{A} \in \mathbb{P}^{m_1 \times p_1}$ and $\mathbf{B} \in \mathbb{P}^{m_2 \times p_2}$ denote transformations on the columns and rows of $\mathbf{X}$. We can rewrite this relation as

$$\mathbf{y} = (\mathbf{B} \otimes \mathbf{A})\,\mathbf{x}, \tag{1}$$

where $\mathbf{y}$ and $\mathbf{x}$ are vectorized versions of $\mathbf{Y}$ and $\mathbf{X}$, and $\otimes$ denotes the matrix Kronecker product [6]. Thus, the DL problem for second-order tensors can be reformulated in terms of (1), where the goal is to find the *Kronecker-structured* (KS) dictionary underlying the data using training samples $\{\mathbf{Y}_i = \mathbf{A}\mathbf{X}_i\mathbf{B}^\top\}_i$. In this work, we take a first step towards understanding the advantages of learning KS dictionaries from second-order tensor data by deriving conditions on the dictionary and coefficient distribution for asymptotic recovery of KS dictionaries from noisy training samples.

A. Relationship to Previous Work

In terms of relation with prior work, there have been several works that develop algorithms to learn KS dictionaries for 2nd-order [7]–[9] and 3rd-order tensors [10] based on the Tucker decomposition of tensors [11]. Although these works demonstrate the enhanced performance of KS-DL compared to traditional vectorized DL schemes, they only focus on the computational aspects of KS-DL and do not provide an understanding of the fundamental advantages and limits associated with learning structured dictionaries for tensor data. There are several works in the literature that study the unstructured dictionary identifiability problem. In Jung et al. [12], minimax lower bounds for dictionary reconstruction error (in the Frobenius norm) are provided. These bounds show that the number of samples for reliable reconstruction (up to a prescribed mean squared error) of a $m \times p$ dictionary within its local neighborhood is on the order of $N = \Omega(mp^2)$. A competing upper bound for the sample complexity of the DL problem is obtained in [13], where it is shown that $N = \Omega(mp^3)$ samples is sufficient to guarantee (with high probability) the existence of a local minimum of the DL cost function within a neighborhood of the true reference dictionary.

In our previous works, we have obtained lower bounds on the minimax risk of KS-DL for 2nd-order [14] and Kth-order tensors [15], [16] and showed that the necessary number of samples for reconstruction of the true KS dictionary within its local neighborhood up to a given estimation error scales with the sum of the product of the dimensions of the coordinate dictionaries, i.e., $N = \Omega(p\sum_{k=1}^{K} m_k p_k)$, compared to $N = \Omega(p\prod_{k=1}^{K} m_k p_k)$ for vectorized data [12]. Given this significant reduction in the lower bound, we turn our attention to the upper bound for the sample complexity of the KS-DL problem and investigate if similar reductions appear there.

B. Our Contributions

In this paper, we study the dictionary identification problem for KS dictionaries given by the Kronecker product of two coordinate dictionaries and derive conditions that ensure the existence of a local minimum of the KS-DL objective function within a small neighborhood of the underlying coordinate dictionaries. Our focus in the analysis is on the asymptotic KS-DL objective function. Our results can lead to an understanding of the finite sample size case, and suggest that KS dictionary

The work of the authors was supported in part by the National Science Foundation under awards CCF-1525276 and CCF-1453073, and by the Army Research Office under award W911NF-17-1-0546.

identification can be done with smaller sample complexity compared to traditional DL identification.

To the best of our knowledge, this is the first work presenting KS-DL identification results. Our proof techniques follow a similar approach as in [13] with the following key distinctions: 1) we assume the reference dictionary underlying the data and the recovered dictionary belong to the class of KS dictionaries, 2) we assume dictionary coefficient vectors follow the *separable sparsity* model that requires non-zero coefficients to be grouped in blocks [15],[1] and 3) we derive conditions that ensure existence of a local minimum within small neighborhoods of the reference coordinate dictionaries.

C. Notation Convention

Bold upper-case and lower-case letters are used to denote matrices and vectors, respectively. Lower-case letters denote scalars. The i-th element of $\mathbf{v}$ is denoted by v_i. The elements of the sign vector of $\mathbf{v}$, denoted as $\operatorname{sign}(\mathbf{v})$, are equal to $\operatorname{sign}(v_i) = v_i/|v_i|$. The k-th column of $\mathbf{X}$ is denoted by $\mathbf{x}_k$ and $\mathbf{X}_\mathcal{I}$ denotes the matrix consisting of the columns of $\mathbf{X}$ with indices $\mathcal{I}$. We use $|\mathcal{I}|$ for the cardinality of the set $\mathcal{I}$. Sometimes we use matrices indexed by numbers, such as $\mathbf{X}_1$, in which case the second index denotes the column index. Norms are given by subscripts, so $\|\mathbf{v}\|_0$, $\|\mathbf{v}\|_1$, and $\|\mathbf{v}\|_2$ are the ℓ_0, ℓ_1, and ℓ_2 norms of $\mathbf{v}$, while $\|\mathbf{X}\|_2$ and $\|\mathbf{X}\|_F$ are the spectral and Frobenius norms of $\mathbf{X}$. We use $[K]$ to denote $\{1, 2, \dots, K\}$. For matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, we define their distance to be $\|\mathbf{X}_1 - \mathbf{X}_2\|_F$. For $\mathbf{X}^0$ belonging to the set $\mathcal{X}$, we define $\mathcal{S}_r(\mathbf{X}^0) \triangleq \left\{\mathbf{X} \in \mathcal{X} : \|\mathbf{X} - \mathbf{X}^0\|_F = r\right\}$, $\mathcal{B}_r(\mathbf{X}^0) \triangleq \left\{\mathbf{X} \in \mathcal{X} : \|\mathbf{X} - \mathbf{X}^0\|_F < r\right\}$, and $\bar{\mathcal{B}}_r(\mathbf{X}^0) \triangleq \left\{\mathbf{X} \in \mathcal{X} : \|\mathbf{X} - \mathbf{X}^0\|_F \le r\right\}$. We use the standard "big-$\mathcal{O}$" notation for asymptotic scaling. We denote $\Delta f(\mathbf{X}_1; \mathbf{X}_2) \triangleq f(\mathbf{X}_1) - f(\mathbf{X}_2)$. We define $\mathbf{H}_\mathbf{X} \triangleq (\mathbf{X}^\top \mathbf{X})^{-1}$, $\mathbf{X}^+ \triangleq \mathbf{H}_\mathbf{X} \mathbf{X}^\top$, and $\mathbf{P}_\mathbf{X} \triangleq \mathbf{X}\mathbf{X}^+$.

II. System Model

We assume observations $\mathbf{y} \in \mathbb{R}^m$ are generated according to a *reference dictionary* $\mathbf{D}^0 = \mathbf{D}_1^0 \otimes \mathbf{D}_2^0$:

$$\mathbf{y} = \left(\mathbf{D}_1^0 \otimes \mathbf{D}_2^0\right) \mathbf{x} + \mathbf{n}, \quad \|\mathbf{x}\|_0 \le s, \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^p$ denotes the sparse generating coefficient vector and $\mathbf{n} \in \mathbb{R}^m$ denotes the underlying noise vector. Here, $\mathbf{D}_k^0 \in \mathcal{D}_k = \{\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}, \|\mathbf{d}_{k,j}\|_2 = 1, \forall j \in [p_k]\}$ for $k \in \{1, 2\}$, and $p_1 p_2 = p$ and $m_1 m_2 = m$. In this work, we study the asymptotic KS-DL objective function. That is, the goal is to recover the underlying KS dictionary by solving the following regularized stochastic optimization program:

$$\min_{\substack{\mathbf{D}_1 \in \mathcal{D}_1 \\ \mathbf{D}_2 \in \mathcal{D}_2}} \mathbb{E}\left\{\inf_{\mathbf{x}' \in \mathbb{R}^p} \frac{1}{2}\left\|\mathbf{y} - (\mathbf{D}_1 \otimes \mathbf{D}_2)\mathbf{x}'\right\|_2^2 + \lambda\|\mathbf{x}'\|_1\right\}, \tag{3}$$

where λ is a regularization parameter and we have replaced the ℓ_0 norm with its convex relaxation, the ℓ_1 norm. We denote the expectation in (3) as $f_\mathbb{P}(\mathbf{D}_1, \mathbf{D}_2)$ and the expression inside the expectation as $f_\mathbf{y}(\mathbf{D}_1, \mathbf{D}_2)$. Our main goal in this paper is to derive conditions that ensure $f_\mathbb{P}(\mathbf{D}_1, \mathbf{D}_2)$ has a local minimum within some neighborhoods of coordinate reference dictionaries $\mathbf{D}_1^0$ and $\mathbf{D}_2^0$.

Coefficient distribution: We assume the coefficient tensor $\mathbf{X}$ follows the *"separable sparsity"* model. Specifically, we sample s_k elements uniformly at random from $[p_k]$, for $k \in \{1, 2\}$. Then, the random support of $\mathbf{x} = \operatorname{vec}(\mathbf{X})$ is $\{\mathcal{J} \subseteq [p], |\mathcal{J}| = s\}$ that is associated with $\{\mathcal{J}_1 \times \mathcal{J}_2 : \mathcal{J}_k \subseteq [p_k], |\mathcal{J}_k| = s_k, k \in \{1, 2\}\}$ via lexicographic indexing and $s = s_1 s_2$. We make the same assumptions as assumptions A and B in [13] and we borrow much of the notation from Gribonval et al. [13]. Some major ones include: $\mathbf{x}$ and $\mathbf{l} = \operatorname{sign}(\mathbf{x})$ are not correlated, the coefficient vectors are bounded, i.e., $\|\mathbf{x}\|_2 \le M_x$, and the nonzero entries of $\mathbf{x}$ have a minimum magnitude, i.e., $\min_{j \in \mathcal{J}} |x_j| \ge x_{\min}$. We also define $\kappa_x \triangleq \mathbb{E}\{|x|\} / \sqrt{\mathbb{E}\{x^2\}}$ as a measure of flatness of $\mathbf{x}$ ($\kappa_x \le 1$ with $\kappa_x = 1$ when all nonzero coefficients are equal) [13].

Noise distribution: We assume additive white bounded noise, i.e., $\|\mathbf{n}\|_2 \le M_n$.

We also use the following definitions throughout the paper: $\delta_k(\mathbf{D})$ denotes the *restricted isometry property* (RIP) constant of order s for matrix $\mathbf{D}$ [19]. For a matrix $\mathbf{D}$, we define its *cumulative coherence* as

$$\mu_s(\mathbf{D}) \triangleq \sup_{|\mathcal{J}| \le s} \sup_{j \notin \mathcal{J}} \|\mathbf{D}_\mathcal{J}^\top \mathbf{d}_j\|_1. \tag{4}$$

Note that for $s = 1$, the cumulative coherence is equivalent to the worst-case coherence [13].

III. Dictionary Identifiability Result

In this section, we provide a variant of [13, Theorem 1] for the KS-DL objective function in (3).

Theorem 1. *Suppose the observations are generated according to* (2) *and the dictionary coefficients follow the separable sparsity model of Section II. Further, assume the following conditions are satisfied:*

$$\max_{k \in \{1,2\}} \left\{\mu_{s_k}(\mathbf{D}_k^0)\right\} \le \frac{1}{4}, \quad s_k \le \frac{p_k}{8\left(\|\mathbf{D}_k^0\|_2 + 1\right)^2}, \tag{5}$$

and $\max\{C_{1,\min}, C_{2,\min}\} < C_{\max}$*, where*

$$\begin{aligned} C_{k,\min} &\triangleq \left(\frac{48}{\sqrt{1.5}}\right) \kappa_x^2 \frac{s_k}{p_k} \left\|\mathbf{D}_k^{0\top} \mathbf{D}_k^0 - \mathbf{I}\right\|_F \left(\|\mathbf{D}_k^0\|_2 + 1\right), \\ C_{\max} &\triangleq \frac{1}{6\sqrt{1.5}} \frac{\mathbb{E}\{|x|\}}{M_x}(1 - 2\mu_s(\mathbf{D}^0)). \end{aligned} \tag{6}$$

Then, the map $(\mathbf{D}_1, \mathbf{D}_2) \in (\mathcal{D}_1, \mathcal{D}_2) \to f_\mathbb{P}(\mathbf{D}_1, \mathbf{D}_2)$ *admits a local minimum* $\widehat{\mathbf{D}} = \widehat{\mathbf{D}}_1 \otimes \widehat{\mathbf{D}}_2$ *such that* $\widehat{\mathbf{D}}_k \in \mathcal{B}_{r_k}(\mathbf{D}_k^0)$*,* $k \in \{1, 2\}$*, for any* $r_k > 0$ *as long as*

$$\lambda \le \frac{x_{\min}}{8\sqrt{1.5}}, \tag{7}$$

$$\lambda C_{k,\min} < \mathbb{E}\{|x|\} r_k < \lambda C_{\max}, \ k \in \{1, 2\}, \tag{8}$$

and

$$M_n < 3\sqrt{1.5} M_x \left(2\frac{\lambda C_{\max}}{\mathbb{E}\{|x|\}} - (r_1 + r_2)\right). \tag{9}$$

The proof of Theorem 1 is provided in Section IV. Here, we provide a brief outline of the formal proof.

[1] This model arises in processing of images and video sequences [17], [18].

For given radii $0 < r_k < 2\sqrt{p_k}, k \in \{1, 2\}$, the spheres $\mathcal{S}_{r_k}(\mathbf{D}_k^0)$ are non-empty.[2] We derive conditions on the coefficients, underlying coordinate dictionaries, noise energy, and r_k's such that

$$\Delta f_{\mathbb{P}}(r_1, r_2) \triangleq \inf_{\mathbf{D}_k \in \mathcal{S}_{r_k}(\mathbf{D}_k^0)} \Delta f_{\mathbb{P}} \left((\mathbf{D}_1, \mathbf{D}_2); (\mathbf{D}_1^0, \mathbf{D}_2^0) \right) > 0.$$

Moreover, the mapping $(\mathbf{D}_1, \mathbf{D}_2) \to f_{\mathbb{P}}(\mathbf{D}_1, \mathbf{D}_2)$ is continuous w.r.t. the Frobenius norm $\|\mathbf{D}_k - \mathbf{D}'_k\|_F$ on all $\mathbf{D}_k, \mathbf{D}'_k \in \mathbb{R}^{m_k \times p_k}, k \in \{1, 2\}$. Hence, it is also continuous on compact constraint sets $\mathcal{D}_k$'s. The compactness of closed balls $\bar{\mathcal{B}}_{r_k}(\mathbf{D}_k^0)$ and the continuity of the mapping $(\mathbf{D}_1, \mathbf{D}_2) \to f_{\mathbb{P}}(\mathbf{D}_1, \mathbf{D}_2)$ imply the existence of a local minimum $(\widehat{\mathbf{D}}_1, \widehat{\mathbf{D}}_2)$ in open balls, $\mathcal{B}_{r_k}(\mathbf{D}_k^0)$'s, $k \in \{1, 2\}$.

To find conditions that ensure $\Delta f_{\mathbb{P}}(r_1, r_2) > 0$, we take the next steps: given coefficients that follow the separable sparsity model, we can decompose any $\mathbf{D}_{\mathcal{J}}, |\mathcal{J}| = s$, as

$$\mathbf{D}_{\mathcal{J}} = \mathbf{D}_{1,\mathcal{J}_1} \otimes \mathbf{D}_{2,\mathcal{J}_2}, \tag{10}$$

where $|\mathcal{J}_k| = s_k$ and $\text{rank}(\mathbf{D}_{k,\mathcal{J}_k}) = s_k$ for $k \in \{1, 2\}$. Given generating $\mathbf{l} = \text{sign}(\mathbf{x})$, we obtain $\widehat{\mathbf{x}}_{\mathbf{y}}((\mathbf{D}_1, \mathbf{D}_2)|\mathbf{l})$ by solving $f_{\mathbf{y}}(\mathbf{D}_1, \mathbf{D}_2)$ conditioned on $\mathbf{l}$, hence eliminating the dependency of $f_{\mathbb{P}}(\mathbf{D}_1, \mathbf{D}_2)$ on $\inf_{\mathbf{x}'}$ by finding a closed-form expression for $f_{\mathbb{P}}(\mathbf{D}_1, \mathbf{D}_2)$ given $\mathbf{l}$, which we denote as $\phi_{\mathbb{P}}\left((\mathbf{D}_1, \mathbf{D}_2)|\mathbf{l}\right)$. Then, assuming $\text{sign}(\widehat{\mathbf{x}}_{\mathbf{y}}((\mathbf{D}_1, \mathbf{D}_2)|\mathbf{l})) = \widehat{\mathbf{l}}$ is equal to $\mathbf{l}$ and using (10), we expand $\Delta\phi_{\mathbb{P}}\left((\mathbf{D}_1, \mathbf{D}_2); (\mathbf{D}_1^0, \mathbf{D}_2^0)|\mathbf{l}\right)$ and separate the terms that depend on each radius $r_k = \|\mathbf{D}_k - \mathbf{D}_k^0\|_F$ to obtain conditions for sparsity levels $s_k, k \in \{1, 2\}$, and coordinate dictionaries such that $\Delta\phi_{\mathbb{P}}\left((\mathbf{D}_1, \mathbf{D}_2); (\mathbf{D}_1^0, \mathbf{D}_2^0)|\mathbf{l}\right) > 0$. Finally, we derive conditions on noise, coordinate dictionary coherences and r_k's that ensure $\Delta f_{\mathbb{P}}\left((\mathbf{D}_1, \mathbf{D}_2); (\mathbf{D}_1^0, \mathbf{D}_2^0)\right) = \Delta\phi_{\mathbb{P}}\left((\mathbf{D}_1, \mathbf{D}_2); (\mathbf{D}_1^0, \mathbf{D}_2^0)|\mathbf{l}\right)$.

Remark 1. The key assumption in the proof of Theorem 1 is expanding $\mathbf{D}_{\mathcal{J}}$ according to (10). This is a consequence of the separable sparsity model for dictionary coefficients. For a detailed discussion on the case where sparse coefficients are drawn uniformly at random (which differs from the separable sparsity model), we refer readers to our earlier work [15].

IV. Proof of Theorem 1

The proof of Theorem 1 relies on the following propositions and lemmas.

Proposition 1. *Suppose the following inequalities hold for* $k \in \{1, 2\}$*:*

$$\max_k \left\{\delta_{s_k}(\mathbf{D}_k^0)\right\} \leq \frac{1}{4} \quad \textit{and} \quad s_k \leq \frac{p_k}{8(\|\mathbf{D}_k^0\|_2 + 1)^2}. \tag{11}$$

Then, for

$$\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\left\{|x|\right\}} \leq \frac{1}{8\sqrt{1.5}}, \tag{12}$$

any $r_k \leq 0.15$*, and for all* $\mathbf{D}_k \in \mathcal{S}_{r_k}(\mathbf{D}_k^0)$*, we have :*

$$\Delta\phi_{\mathbb{P}}\left((\mathbf{D}_1, \mathbf{D}_2); (\mathbf{D}_1^0, \mathbf{D}_2^0)|\mathbf{l}\right) \geq \frac{\mathbb{E}\{x^2\}s}{8} \left(\frac{r_1}{p_1}\left(r_1 - r_{1,\min}(\bar{\lambda})\right) + \frac{r_2}{p_2}\left(r_2 - r_{2,\min}(\bar{\lambda})\right) \right),$$

where $r_{k,\min}(\bar{\lambda}) \triangleq \left(1.5 + \frac{8\sqrt{1.5}}{3}\bar{\lambda}\right)\bar{\lambda}C_{k,\min}$ *. In addition, if* $\bar{\lambda} \leq 0.15/\max_k C_{k,\min}$*, then* $r_{k,\min} < 0.15$*. Thus,* $\Delta\phi_{\mathbb{P}}\left((\mathbf{D}_1, \mathbf{D}_2); (\mathbf{D}_1^0, \mathbf{D}_2^0)|\mathbf{l}\right) \geq 0$ *for all* $r_k \in (r_{k,\min}(\bar{\lambda}), 0.15], k \in \{1, 2\}$.

The proof of Proposition 1 relies on the following definition and lemmas as well as Lemmas 4–7,15, and 16 in [13].

Definition 1. Given $(\mathbf{D}_1, \mathbf{D}_2)$ and $(\mathbf{D}_1^0, \mathbf{D}_2^0)$, we have

$$\begin{aligned}
&(\mathbf{D}_1 \otimes \mathbf{D}_2) - (\mathbf{D}_1^0 \otimes \mathbf{D}_2^0) \\
&\quad = (\mathbf{D}_1 - \mathbf{D}_1^0) \otimes \mathbf{D}_2 + \mathbf{D}_1^0 \otimes (\mathbf{D}_2 - \mathbf{D}_2^0) \\
&\quad = (\mathbf{D}_1 - \mathbf{D}_1^0) \otimes \mathbf{D}_2^0 + \mathbf{D}_1 \otimes (\mathbf{D}_2 - \mathbf{D}_2^0) \\
&\quad \triangleq (\mathbf{D}_1 - \mathbf{D}_1^0) \otimes \widetilde{\mathbf{D}}_2 + \widetilde{\mathbf{D}}_1 \otimes (\mathbf{D}_2 - \mathbf{D}_2^0),
\end{aligned} \tag{13}$$

where without loss of generality, we have defined $\widetilde{\mathbf{D}}_k$ to be equal to either $\mathbf{D}_k^0$ or $\mathbf{D}_k$.

Lemma 1. *Let* $\mathbf{l} \in \{-1, 0, 1\}^p$ *be an arbitrary sign vector and* $\mathcal{J} = \mathcal{J}(\mathbf{l})$ *be its support. Define*

$$\phi_{\mathbf{y}}((\mathbf{D}_1, \mathbf{D}_2)|\mathbf{l}) = \inf_{\substack{\mathbf{x} \in \mathbb{R}^p \\ \text{supp}(\mathbf{x}) \subset \mathcal{J}}} \frac{1}{2}\left\|\mathbf{y} - (\mathbf{D}_1 \otimes \mathbf{D}_2)\,\mathbf{x}\right\|_2^2 + \lambda \mathbf{l}^\top \mathbf{x}.$$

Then, if $\mathbf{D}_{k,\mathcal{J}_k}^\top \mathbf{D}_{k,\mathcal{J}_k}$ *is invertible for* $k \in \{1, 2\}$, $\phi_{\mathbf{y}}((\mathbf{D}_1, \mathbf{D}_2)|\mathbf{l})$ *can be expressed in closed form:*

$$\begin{aligned}
\phi_{\mathbf{y}}((\mathbf{D}_1, \mathbf{D}_2)|\mathbf{l}) &= \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\mathbf{y}^\top \left(\mathbf{P}_{\mathbf{D}_{1,\mathcal{J}_1}} \otimes \mathbf{P}_{\mathbf{D}_{2,\mathcal{J}_2}}\right)\mathbf{y} \\
&+ \lambda \mathbf{l}_{\mathcal{J}}^\top \left(\mathbf{D}_{1,\mathcal{J}_1}^+ \otimes \mathbf{D}_{2,\mathcal{J}_2}^+\right)\mathbf{y} - \frac{\lambda^2}{2}\mathbf{l}_{\mathcal{J}}^\top \left(\mathbf{H}_{\mathbf{D}_{1,\mathcal{J}_1}} \otimes \mathbf{H}_{\mathbf{D}_{2,\mathcal{J}_2}}\right)\mathbf{l}_{\mathcal{J}}.
\end{aligned}$$

Lemma 2. *Assume* $\max\left\{\delta_{s_k}(\mathbf{D}_k^0), \delta_{s_k}(\mathbf{D}_k)\right\} < 1$ *for* $k \in \{1, 2\}$*. For* $\phi_{\mathbb{P}}\left((\mathbf{D}_1, \mathbf{D}_2)|\mathbf{l}\right) \triangleq \mathbb{E}\left\{\phi_{\mathbf{y}}((\mathbf{D}_1, \mathbf{D}_2)|\mathbf{l})\right\}$*, we have*

$$\begin{aligned}
&\Delta\phi_{\mathbb{P}}\left((\mathbf{D}_1, \mathbf{D}_2); (\mathbf{D}_1^0, \mathbf{D}_2^0)|\mathbf{l}\right) = \frac{\mathbb{E}\{x^2\}}{2} \\
&\Big[\mathbb{E}\left\{\text{Tr}\left[{\mathbf{D}_1^0}^\top \mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}} \mathbf{D}_1^0\right]\right\} \mathbb{E}\left\{\text{Tr}\left[{\mathbf{D}_2^0}^\top (\mathbf{I} - \mathbf{P}_{\mathbf{D}_{2,\mathcal{J}_2}})\mathbf{D}_2^0\right]\right\} \\
&+\mathbb{E}\left\{\text{Tr}\left[{\mathbf{D}_1^0}^\top (\mathbf{I} - \mathbf{P}_{\mathbf{D}_{1,\mathcal{J}_1}})\mathbf{D}_1^0\right]\right\} \mathbb{E}\left\{\text{Tr}\left[{\mathbf{D}_2^0}^\top \mathbf{P}_{\widetilde{\mathbf{D}}_{2,\mathcal{J}_2}} \mathbf{D}_2^0\right]\right\}\Big] \\
&- \lambda\mathbb{E}\{|x|\}\Big[\mathbb{E}\left\{\text{Tr}\left[\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}^+ \mathbf{D}_1^0\right]\right\} \mathbb{E}\left\{\text{Tr}\left[\mathbf{I} - \mathbf{D}_{2,\mathcal{J}_2}^+ \mathbf{D}_2^0\right]\right\} \\
&+\mathbb{E}\left\{\text{Tr}\left[\mathbf{I} - \mathbf{D}_{1,\mathcal{J}_1}^+ \mathbf{D}_1^0\right]\right\} \mathbb{E}\left\{\text{Tr}\left[\widetilde{\mathbf{D}}_{2,\mathcal{J}_2}^+ \mathbf{D}_2^0\right]\right\}\Big] \\
&+ \frac{\lambda^2}{2}\Big[\mathbb{E}\left\{\text{Tr}\left[\mathbf{H}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\right]\right\} \mathbb{E}\left\{\text{Tr}\left[\mathbf{H}_{\mathbf{D}_{2,\mathcal{J}_2}^0} - \mathbf{H}_{\mathbf{D}_{2,\mathcal{J}_2}}\right]\right\} \\
&+\mathbb{E}\left\{\text{Tr}\left[\mathbf{H}_{\mathbf{D}_{1,\mathcal{J}_1}^0} - \mathbf{H}_{\mathbf{D}_{1,\mathcal{J}_1}}\right]\right\} \mathbb{E}\left\{\text{Tr}\left[\mathbf{H}_{\widetilde{\mathbf{D}}_{2,\mathcal{J}_2}}\right]\right\}\Big].
\end{aligned} \tag{14}$$

Lemma 3. *For any* $\mathcal{J}_k \subset [p_k], |\mathcal{J}_k| = s_k$*, the following relations hold:*

$$\begin{aligned}
\left\|\mathbf{D}_{k,\mathcal{J}_k}\right\|_2 &\leq \sqrt{1 + \delta_{s_k}(\mathbf{D}_k)}, \\
\left\|\mathbf{D}_{k,\mathcal{J}_k}^\top\right\|_2 &\leq \sqrt{1 + \mu_{s_k - 1}(\mathbf{D}_k)}, \\
\delta_{s_k}(\mathbf{D}_k) &\leq \mu_{s_k - 1}(\mathbf{D}_k).
\end{aligned} \tag{15}$$

Lemma 4. *Suppose that* $(\mathbf{D}_1, \mathbf{D}_2)$ *and* $(\mathbf{D}_1^0, \mathbf{D}_2^0)$ *are such that*

$$\begin{aligned}
A_k &\geq \max\left\{\left\|\mathbf{D}_k^\top \mathbf{D}_k - \mathbf{I}\right\|_F, \left\|{\mathbf{D}_k^0}^\top \mathbf{D}_k^0 - \mathbf{I}\right\|_F\right\}, \\
B_k &\geq \max\left\{\|\mathbf{D}_k\|_2, \|\mathbf{D}_k^0\|_2\right\}, \\
\delta_k &\geq \max\left\{\delta_{s_k}(\mathbf{D}_k), \delta_{s_k}(\mathbf{D}_k^0)\right\}.
\end{aligned} \tag{16}$$

[2] This follows from the construction of dictionary classes, $\mathcal{D}_k$'s.

Then, we have

$$\Delta\phi_{\mathbb{P}}\left((\mathbf{D}_1,\mathbf{D}_2);(\mathbf{D}_1^0,\mathbf{D}_2^0)|\mathbf{l}\right) \geq \frac{\mathbb{E}\{x^2\}}{2}\sum_{k=1}^{2}\frac{s}{p_k}\|\mathbf{D}_k-\mathbf{D}_k^0\|_F$$
$$\Bigg[\|\mathbf{D}_k-\mathbf{D}_k^0\|_F\left(1-\frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k}-\bar{\lambda}\kappa_x^2\left(\sqrt{\frac{1+\delta_{k'}}{1-\delta_{k'}}}\right)\right)$$
$$-\left(\sqrt{\frac{1+\delta_{k'}}{1-\delta_{k'}}}+\frac{2\bar{\lambda}}{(1-\delta_1)(1-\delta_2)}\right)\bar{\lambda}\kappa_x^2\frac{s_k}{p_k}\frac{2A_kB_k}{1-\delta_k}\Bigg],$$

where $k'=1$ if $k=2$ and $k'=2$ if $k=1$.

Proposition 1 shows $\Delta\phi_{\mathbb{P}}((\mathbf{D}_1,\mathbf{D}_2);(\mathbf{D}_1^0,\mathbf{D}_2^0)|\mathbf{l}) \geq 0$. However, given $\widehat{\mathbf{x}}_{\mathbf{y}}((\mathbf{D}_1,\mathbf{D}_2)|\mathbf{l})$ the solution of $\phi_{\mathbf{y}}((\mathbf{D}_1,\mathbf{D}_2)|\mathbf{l})$, $\widehat{\mathbf{l}}=\operatorname{sign}\left(\widehat{\mathbf{x}}_{\mathbf{y}}((\mathbf{D}_1,\mathbf{D}_2)|\mathbf{l})\right)$ is not necessarily equal to $\mathbf{l}$. We derive conditions that ensure $\widehat{\mathbf{x}}_{\mathbf{y}}((\mathbf{D}_1,\mathbf{D}_2)|\mathbf{l})$ is almost surely the unique minimizer of $f_{\mathbf{y}}(\mathbf{D}_1,\mathbf{D}_2)$ and $\widehat{\mathbf{l}}=\mathbf{l}$. We introduce the following proposition for this purpose.

Proposition 2. *Let the reference coordinate dictionaries $\mathbf{D}_1^0$ and $\mathbf{D}_2^0$ satisfy:*

$$\max_k\{\delta_{s_k}(\mathbf{D}_k^0)\} \leq \frac{1}{4} \quad \textit{and} \quad \max_k\{\mu_{s_k}(\mathbf{D}_k^0)\} \leq \frac{1}{2}. \tag{17}$$

Suppose $\bar{\lambda} \leq \dfrac{x_{\min}}{2\mathbb{E}\{|x|\}}$ and $\max\{r_1,r_2\} \leq \bar{\lambda}C_{\max}$. If the noise level satisfies

$$M_n < 3\sqrt{1.5}M_x\left(2\bar{\lambda}C_{\max}-(r_1+r_2)\right), \tag{18}$$

then, for any $(\mathbf{D}_1,\mathbf{D}_2)$ such that $\mathbf{D}_k \in \mathcal{S}_{r_k}(\mathbf{D}_k^0)$, for $k \in \{1,2\}$, $\widehat{\mathbf{x}}_{\mathbf{y}}((\mathbf{D}_1,\mathbf{D}_2)|\mathbf{l})$ is almost surely the minimizer of $\mathbf{x} \to \frac{1}{2}\left\|\mathbf{y}-(\mathbf{D}_1\otimes\mathbf{D}_2)\,\mathbf{x}\right\|_2^2+\lambda\|\mathbf{x}\|_1$, $\widehat{\mathbf{l}}=\mathbf{l}$ and

$$\Delta\phi_{\mathbb{P}}\left((\mathbf{D}_1,\mathbf{D}_2),(\mathbf{D}_1^0,\mathbf{D}_2^0)|\mathbf{l}\right)=\Delta f_{\mathbb{P}}\left((\mathbf{D}_1,\mathbf{D}_2),(\mathbf{D}_1^0,\mathbf{D}_2^0)\right).$$

The proof of Proposition 2 relies on the following lemma and Lemmas 10–13 in [13]:

Lemma 5. *For any $\mathbf{D}^0=\mathbf{D}_1^0\otimes\mathbf{D}_2^0$ and $\mathbf{D}=\mathbf{D}_1\otimes\mathbf{D}_2$ such that $\mathbf{D}_k \in \bar{\mathcal{B}}_{r_k}(\mathbf{D}_k^0)$, for $k \in \{1,2\}$, suppose the following inequalities are satisfied:*

$$\max_k\{\mu_{s_k}(\mathbf{D}_k^0)\} \leq \frac{1}{2} \quad \textit{and} \quad \max_k\{\delta_{s_k}(\mathbf{D}_k^0)\} \leq \frac{1}{4}. \tag{19}$$

Then, we have

$$\mu_s(\mathbf{D}) \leq \mu_s(\mathbf{D}^0)+2\sqrt{1.5s}\,(r_1+r_2). \tag{20}$$

Proof of Theorem 1: The assumptions in (5) ensure that the conditions in (11) are satisfied for Proposition 1 and the conditions in (17) hold for Proposition 2 . The assumption in (6) implies

$$\frac{\mathbb{E}\{x^2\}}{M_x\mathbb{E}\{|x|\}} > \left(\frac{288}{(1-2\mu_s(\mathbf{D}^0))}\right)$$
$$\max_{k=\{1,2\}}\left\{\frac{s_k}{p_k}\left\|{\mathbf{D}_k^0}^\top\mathbf{D}_k^0-\mathbf{I}\right\|_F\left(\left\|\mathbf{D}_k^0\right\|_2+1\right)\right\}. \tag{21}$$

Equation (21) and (7) ensure that the condition in (12) is satisfied for Proposition 1 and $\bar{\lambda} \leq \dfrac{x_{\min}}{2\mathbb{E}\{|x|\}}$ holds for Proposition 2. Hence, according to Proposition 1, $\Delta\phi_{\mathbb{P}}\left((\mathbf{D}_1,\mathbf{D}_2);(\mathbf{D}_1^0,\mathbf{D}_2^0)|\mathbf{l}\right) \geq 0$ for all $r_k \in (\bar{\lambda}C_{k,\min},0.15], k \in \{1,2\}$. Finally, according to Proposition 2, the assumption in (9) implies $\Delta\phi_{\mathbb{P}}\left((\mathbf{D}_1,\mathbf{D}_2);(\mathbf{D}_1^0,\mathbf{D}_2^0)|\mathbf{l}\right) = \Delta f_{\mathbb{P}}\left((\mathbf{D}_1,\mathbf{D}_2);(\mathbf{D}_1^0,\mathbf{D}_2^0)\right)$ for all $r_k \leq \bar{\lambda}C_{\max}, k \in \{1,2\}$. Furthermore, the assumption in (7) implies $C_{\max}\bar{\lambda} \leq 0.15$. Consequently, for any $r_k > 0$, $k \in \{1,2\}$ satisfying conditions in (8), $(\mathbf{D}_1,\mathbf{D}_2) \in (\mathcal{D}_1,\mathcal{D}_2) \to f_{\mathbb{P}}(\mathbf{D}_1,\mathbf{D}_2)$ admits a local minimum $\widehat{\mathbf{D}}=\widehat{\mathbf{D}}_1\otimes\widehat{\mathbf{D}}_2$ such that $\widehat{\mathbf{D}}_k \in \mathcal{B}_{r_k}(\mathbf{D}_k^0), k \in \{1,2\}$. ■

V. Discussion

In this section, we discuss the implications of Theorem 1. Comparing this result with the results of [13, Theorem 1] for vectorized observations, we see that our result captures the dependence of the local minimum on the coordinate dictionaries and, also, demonstrates that there exists a local minimum of $f_{\mathbb{P}}(\mathbf{D}_1,\mathbf{D}_2)$ that is in a local neighborhood of the coordinate dictionaries. This ensures recovery of coordinate dictionaries (within some local neighborhood of true coordinate dictionaries), as opposed to KS dictionary recovery [13], suggesting that we can obtain smaller sample complexity results for KS-DL that depend on the neighborhood radii around coordinate dictionaries and their dimensions.

Comparing our conditions for Theorem 1 with [13, Theorem 1], given coefficients drawn from the separable sparsity model, the sparsity constraints for the coordinate dictionaries in (5) translate into

$$\frac{s}{p}=\frac{s_1s_2}{p} \leq \frac{1}{64\left(\|\mathbf{D}_1^0\|_2+1\right)^2\left(\|\mathbf{D}_2^0\|_2+1\right)^2}. \tag{22}$$

Therefore, we have $\mathcal{O}\left(\dfrac{s}{p}\right) = \dfrac{1}{\|\mathbf{D}_1^0\|_2^2\,\|\mathbf{D}_2^0\|_2^2} = \dfrac{1}{\|\mathbf{D}^0\|_2^2}$. Using the fact that $\left\|\mathbf{D}^0\right\|_2 \geq \|\mathbf{D}^0\|_F/\sqrt{m} = \sqrt{p}/\sqrt{m}$, this translates into sparsity order $s=\mathcal{O}(m)$. This scaling is similar to the scaling of the sparsity level in [13]. Moreover, looking at the left hand side of the condition in (21), it is less than 1. According to the Welch bound [20], we have

$$\left\|{\mathbf{D}_k^0}^\top\mathbf{D}_k^0-\mathbf{I}\right\|_F \geq \sqrt{\frac{p_k(p_k-m_k)}{m_k}}. \tag{23}$$

The fact that $\left\|\mathbf{D}_k^0\right\|_2 \geq \sqrt{p_k}/\sqrt{m_k}$ and the assumption $\mu_{s_k}(\mathbf{D}_k^0) \leq 1/4$ imply that the right hand side of (21) is lower bounded by $\Omega\left(\max_k s_k\sqrt{(p_k-m_k)/m_k^2}\right)$. Therefore, Theorem 1 applies to coordinate dictionaries with dimensions $p_k \leq m_k^2$ and subsequently, KS-dictionary with $p \leq m^2$. This is in line with the scaling results in [13] for vectorized signals.

VI. Conclusion

In this work, we addressed the KS-DL identification problem and obtained conditions on the dictionary coefficients, noise level, and the underlying KS dictionary that ensure existence of a local minimum of the KS-DL asymptotic objective function within local neighborhoods of the coordinate dictionaries. Future work includes providing sample complexity results for the KS dictionary identification problem, generalizing the analysis to KS-DL for Kth-order tensor data, extension of the results to randomly sparse distributed coefficient models, and providing KS-DL algorithms that achieve the sample complexity scaling.

References

[1] J. C. Harsanyi and C.-I. Chang, "Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 4, pp. 779–785, July 1994. [Online]. Available: http://dx.doi.org/10.1109/36.298007

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[3] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, February 2003. [Online]. Available: http://dx.doi.org/10.1162/089976603762552951

[4] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, November 2006. [Online]. Available: http://dx.doi.org/10.1109/TSP.2006.881199

[5] O. Bryt and M. Elad, "Compression of facial images using the k-svd algorithm," *Journal of Visual Communication and Image Representation*, vol. 19, no. 4, pp. 270–282, May 2008. [Online]. Available: https://dx.doi.org/10.1016/j.jvcir.2008.03.001

[6] C. F. Van Loan, "The ubiquitous Kronecker product," *Journal of Computational and Applied Mathematics*, vol. 123, no. 1, pp. 85–100, November 2000. [Online]. Available: http://dx.doi.org/10.1016/S0377-0427(00)00393-9

[7] S. Hawe, M. Seibert, and M. Kleinsteuber, "Separable dictionary learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 438–445. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2013.63

[8] F. Roemer, G. Del Galdo, and M. Haardt, "Tensor-based algorithms for learning multidimensional separable dictionaries," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 3963–3967. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2014.6854345

[9] C. F. Dantas, M. N. da Costa, and R. da Rocha Lopes, "Learning dictionaries as a sum of kronecker products," *IEEE Sig. Process. Letters*, vol. 24, no. 5, pp. 559–563, 2017. [Online]. Available: http://dx.doi.org/10.1109/LSP.2017.2681159

[10] S. Zubair and W. Wang, "Tensor dictionary learning with sparse Tucker decomposition," in *Proc. IEEE 18th Int. Conf. Digital Signal Process. (DSP)*, July 2013, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICDSP.2013.6622725

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

[12] A. Jung, Y. C. Eldar, and N. Görtz, "On the minimax risk of dictionary learning," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1501–1515, March 2015. [Online]. Available: http://dx.doi.org/10.1109/TIT.2016.2517006

[13] 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, November 2015. [Online]. Available: http://dx.doi.org/10.1109/TIT.2015.2472522

[14] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds for Kronecker-structured dictionary learning," in *Proc. IEEE Int. Symp. Inf. Theory*, July 2016, pp. 1148–1152. [Online]. Available: http://dx.doi.org/10.1109/ISIT.2016.7541479

[15] ——, "Minimax lower bounds on dictionary learning for tensor data," *arXiv preprint arXiv:1608.02792*, August 2016.

[16] ——, "Sample complexity bounds for dictionary learning of tensor data," in *Proc. 41st IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, March 2017, pp. 4501–4505. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2017.7953008

[17] C. F. Caiafa and A. Cichocki, "Computing sparse representations of multidimensional signals using Kronecker bases," *Neural Computation*, vol. 25, no. 1, pp. 186–220, January 2013. [Online]. Available: http://dx.doi.org/10.1162/NECO_a_00385

[18] ——, "Multidimensional compressed sensing and their applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 6, pp. 355–380, October 2013. [Online]. Available: http://dx.doi.org/10.1002/widm.1108

[19] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008. [Online]. Available: https://doi.org/10.1016/j.crma.2008.03.014

[20] L. R. Welch, "Lower bounds on the maximum cross correlation of signals," *IEEE Transactions on Information Theory*, vol. 20, no. 3, pp. 397–399, 1974. [Online]. Available: http://dx.doi.org/10.1109/TIT.1974.1055219