Minimax Lower Bounds for Kronecker-Structured Dictionary Learning

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

***Abstract*—Dictionary learning is the problem of estimating the collection of atomic elements that provide a sparse representation of measured/collected signals or data. This paper finds fundamental limits on the sample complexity of estimating dictionaries for tensor data by proving a lower bound on the minimax risk. This lower bound depends on the dimensions of the tensor and parameters of the generative model. The focus of this paper is on second-order tensor data, with the underlying dictionaries constructed by taking the Kronecker product of two smaller dictionaries and the observed data generated by sparse linear combinations of dictionary atoms observed through white Gaussian noise. In this regard, the paper provides a general lower bound on the minimax risk and also adapts the proof techniques for equivalent results using sparse and Gaussian coefficient models. The reported results suggest that the sample complexity of dictionary learning for tensor data can be significantly lower than that for unstructured data.**

I. Introduction

Dictionary learning has recently received significant attention due to the increased importance of finding sparse representations of signals/data. In dictionary learning, the goal is to construct an overcomplete basis using input signals such that each signal can be described by a small number of atoms (columns) [1]. Although the existing literature has focused on one-dimensional data, many signals in practice are multi-dimensional and have a tensor structure: examples include 2-dimensional images and 3-dimensional signals produced via magnetic resonance imaging or computed tomography systems. In traditional dictionary learning techniques, multi-dimensional data are processed after vectorizing of signals. This can result in poor sparse representations as the structure of the data is neglected [2].

In this paper we provide fundamental limits on learning dictionaries for multi-dimensional data with tensor structure: we call such dictionaries *Kronecker-structured* (KS). Several algorithms have been proposed to learn KS dictionaries [2]–[7] but there has been little work on the theoretical guarantees of such algorithms. The lower bounds we provide on the minimax risk of learning a KS dictionary give a measure to evaluate the performance of the existing algorithms.

In terms of relation to prior work, theoretical insights into classical dictionary learning techniques [8]–[16] have either focused on achievability of existing algorithms [8]–[14] or lower bounds on minimax risk for one-dimensional data [15], [16]. The former works provide sample complexity results for reliable dictionary estimation based on the appropriate minimization criteria [8]–[14]. Specifically, given a probabilistic model for sparse signals and a finite number of samples, a dictionary is recoverable within some distance of the true dictionary as a local minimum of some minimization criterion [12]–[14]. In contrast, works like Jung et al. [15], [16] provide minimax lower bounds for dictionary learning under several coefficient vector distributions and discuss a regime where the bounds are tight for some signal-to-noise (SNR) values. Particularly, for a dictionary $\mathbf{D} \in \mathbb{R}^{m\times p}$ and neighborhood radius r, they show $N = \mathcal{O}(r^2mp)$ samples suffices for reliable recovery of the dictionary within its local neighborhood.

While our work is related to that of Jung et al. [15], [16], our main contribution is providing lower bounds for the minimax risk of dictionaries consisting of two coordinate dictionaries that sparsely represent 2-dimensional tensor data. The full version of this work generalizes the results to higher-order tensors [17]. The main approach taken in this regard is the well-understood technique of lower bounding the minimax risk in nonparametric estimation by the maximum probability of error in a carefully constructed multiple hypothesis testing problem [18], [19]. As such, our general approach is similar to the vector case [16]. Nonetheless, the major challenge in such minimax risk analyses is the construction of appropriate multiple hypotheses, which are fundamentally different in our problem setup due to the Kronecker structure of the true dictionary. In particular, for a dictionary $\mathbf{D}$ consisting of the Kronecker product of two coordinate dictionaries $\mathbf{A} \in \mathbb{R}^{m_1\times p_1}$ and $\mathbf{B} \in \mathbb{R}^{m_2\times p_2}$, where $m = m_1m_2$ and $p = p_1p_2$, our analysis reduces the sample complexity from $\mathcal{O}(r^2mp)$ for vectorized data [16] to $\mathcal{O}(r^2(m_1p_1+m_2p_2))$. Our results hold even when one of the coordinate dictionaries is not overcomplete (note that both $\mathbf{A}$ and $\mathbf{B}$ cannot be undercomplete, otherwise $\mathbf{D}$ won't be overcomplete). Like previous work [16], our analysis is local and our lower bounds depend on the distribution of multidimensional data. Finally, some of our analysis relies on the availability of side information about the signal samples. This suggests that the lower bounds can be improved by deriving them in the absence of such side information.

Notational Convention: Underlined bold upper-case, bold upper-case, bold lower-case and lower-case letters are used to denote tensors, matrices, vectors, and scalars, respectively. We write $[K]$ for $\{1,\dots,K\}$. The k-th column of a matrix $\mathbf{X}$ is denoted by $\mathbf{x}_k$, while $\mathbf{X}_{\mathcal{I}}$ denotes the matrix consisting

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-14-1-0295.

978-1-5090-1806-2/16/$31.00 ©2016 IEEE

of columns of $\mathbf{X}$ with indices $\mathcal{I}$, $\sum \mathbf{X}$ denotes the sum of all elements of $\mathbf{X}$, and $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. Also, $\|\mathbf{v}\|_0$ and $\|\mathbf{v}\|_2$ denote the ℓ_0 and ℓ_2 norms of the vector $\mathbf{v}$, respectively, while $\|\mathbf{X}\|_2$ and $\|\mathbf{X}\|_F$ denote the spectral and Frobenius norms of $\mathbf{X}$, respectively.

We write $\mathbf{X}_1 \otimes \mathbf{X}_2$ for the *Kronecker product* of two matrices $\mathbf{X}_1 \in \mathbb{R}^{m \times n}$ and $\mathbf{X}_2 \in \mathbb{R}^{p \times q}$: the result is an $mp \times nq$ matrix. Given $\mathbf{X}_1 \in \mathbb{R}^{m \times n}$ and $\mathbf{X}_2 \in \mathbb{R}^{p \times n}$, we write $\mathbf{X}_1 * \mathbf{X}_2$ for their $mp \times n$ *Khatri-Rao product* [20]: this is essentially the column-wise Kronecker product of matrices. Given two matrices of the same dimension $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{m \times n}$, their $m \times n$ *Hadamard product* is denoted by $\mathbf{X}_1 \odot \mathbf{X}_2$, which is the element-wise product of $\mathbf{X}_1$ and $\mathbf{X}_2$. For matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, we define their distance to be $\|\mathbf{X}_1 - \mathbf{X}_2\|_F$. We use $f(\varepsilon) = \mathcal{O}(g(\varepsilon))$ if $\lim_{\varepsilon \to 0} f(\varepsilon)/g(\varepsilon) = c < \infty$ for some constant c.

II. Background and Problem Formulation

In the conventional dictionary learning setup, it is assumed that an observation $\mathbf{y} \in \mathbb{R}^m$ is generated via a fixed dictionary,

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{n}, \tag{1}$$

in which the dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ is an overcomplete basis ($m < p$) with unit-norm columns, $\mathbf{x} \in \mathbb{R}^p$ is the coefficient vector, and $\mathbf{n} \in \mathbb{R}^m$ is the underlying noise vector. In contrast to this conventional setup, our focus in this paper is on second-order tensor data. Consider the 2-dimensional observation $\underline{\mathbf{Y}} \in \mathbb{R}^{m_1 \times m_2}$. Using any separable transform, $\underline{\mathbf{Y}}$ can be written as

$$\underline{\mathbf{Y}} = (\mathbf{T}_1^{-1})^T \underline{\mathbf{X}} \mathbf{T}_2^{-1}, \tag{2}$$

where $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times p_2}$ is the matrix of coefficients and $\mathbf{T}_1 \in \mathbb{R}^{p_1 \times m_1}$ and $\mathbf{T}_2 \in \mathbb{R}^{p_2 \times m_2}$ are non-singular matrices transforming the columns and rows of $\underline{\mathbf{Y}}$, respectively. Defining $\mathbf{A} \triangleq (\mathbf{T}_2^{-1})^T$ and $\mathbf{B} \triangleq (\mathbf{T}_1^{-1})^T$, we can use a property of Kronecker products [21], $\mathrm{vec}(\mathbf{B}\underline{\mathbf{X}}\mathbf{A}^T) = (\mathbf{A} \otimes \mathbf{B})\,\mathrm{vec}(\underline{\mathbf{X}})$, to get the following expression for $\mathbf{y} \triangleq \mathrm{vec}(\underline{\mathbf{Y}})$:

$$\mathbf{y} = (\mathbf{A} \otimes \mathbf{B})\mathbf{x} + \mathbf{n} \tag{3}$$

for coefficient vector $\mathbf{x} \triangleq \mathrm{vec}(\underline{\mathbf{X}}) \in \mathbb{R}^p$, and noise vector $\mathbf{n} \in \mathbb{R}^m$, where $p \triangleq p_1 p_2$ and $m \triangleq m_1 m_2$. In this work, we assume N independent and identically distributed (i.i.d.) noisy observations $\mathbf{y}_k$ that are generated according to the model in (3). Concatenating these observations in $\mathbf{Y} \in \mathbb{R}^{m \times N}$, we have

$$\mathbf{Y} = \mathbf{D}\mathbf{X} + \mathbf{N}, \tag{4}$$

where $\mathbf{D} \triangleq \mathbf{A} \otimes \mathbf{B}$ is the unknown KS dictionary, $\mathbf{X} \in \mathbb{R}^{p \times N}$ is the coefficient matrix which we initially assume to consist of zero-mean random coefficient vectors with known distribution and covariance $\mathbf{\Sigma}_x$, and $\mathbf{N} \in \mathbb{R}^{m \times N}$ is additive white Gaussian noise (AWGN) with zero mean and variance σ^2.

Our main goal in this paper is to derive conditions under which the dictionary $\mathbf{D}$ can possibly be learned from the noisy observations given in (4). In this regard, we assume the true KS dictionary $\mathbf{D}$ consists of unit norm columns and we carry out local analysis. That is, the true KS dictionary $\mathbf{D}$ is assumed to belong to a neighborhood around a fixed (normalized) reference KS dictionary $\mathbf{D}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0$, i.e., $\|\mathbf{a}_{0,j}\|_2 = 1\ \forall j \in [p_1]$, $\|\mathbf{b}_{0,j}\|_2 = 1\ \forall j \in [p_2]$, and $\mathbf{D}_0 \in \mathcal{D}$:

$$\mathcal{D} \triangleq \big\{\mathbf{D}' \in \mathbb{R}^{m \times p} : \|\mathbf{d}'_j\|_2 = 1\ \forall j \in [p],\ \mathbf{D}' = \mathbf{A}' \otimes \mathbf{B}', \\ \mathbf{A}' \in \mathbb{R}^{m_1 \times p_1}, \mathbf{B}' \in \mathbb{R}^{m_2 \times p_2}\big\},\ \text{and} \tag{5}$$

$$\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r) \triangleq \left\{\mathbf{D}' \in \mathcal{D} : \|\mathbf{D}' - \mathbf{D}_0\|_F^2 < r\right\}, \tag{6}$$

where the radius r is known. It is worth noting here that, similar to the analysis for vector data [16], our analysis is applicable to the global KS dictionary learning problem. Finally, some of our analysis in the following also relies on the notion of the *restricted isometry property* (RIP). Specifically, $\mathbf{D}$ satisfies the RIP of order s with constant δ_s if

$$\forall\ s\text{-sparse } \mathbf{x}, (1 - \delta_s)\|\mathbf{x}\|_2^2 \le \|\mathbf{D}\mathbf{x}\|_2^2 \le (1 + \delta_s)\|\mathbf{x}\|_2^2. \tag{7}$$

A. Minimax risk analysis

We are interested in lower bounding the minimax risk for estimating $\mathbf{D}$ based on observations $\mathbf{Y}$, which is defined as the worst-case mean squared error (MSE) that can be obtained by the best KS dictionary estimator $\widehat{\mathbf{D}}(\mathbf{Y})$. That is,

$$\varepsilon^* = \inf_{\widehat{\mathbf{D}}} \sup_{\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r)} \mathbb{E}_{\mathbf{Y}} \left\{\left\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\right\|_F^2\right\}. \tag{8}$$

In order to lower bound this minimax risk ε^*, we resort to the multiple hypothesis testing approach taken in the literature on nonparametric estimation [18], [19]. This approach is equivalent to generating a KS dictionary $\mathbf{D}_l$ uniformly at random from a carefully constructed class $\mathcal{D}_L = \{\mathbf{D}_1, \ldots, \mathbf{D}_L\} \subseteq \mathcal{X}(\mathbf{D}_0, r), L \ge 2$, for a given $(\mathbf{D}_0, r)$. Observations $\mathbf{Y} = \mathbf{D}_l\mathbf{X} + \mathbf{N}$ in this setting can be interpreted as channel outputs that are fed into an estimator that must decode $\mathbf{D}_l$. A lower-bound on the minimax risk in this setting depends not only on problem parameters such as the number of observations N, noise variance σ^2, dimensions of the true KS dictionary, neighborhood radius r, and coefficient distribution, but also on various aspects of the constructed class $\mathcal{D}_L$ [18].

To ensure a tight lower bound, we must construct $\mathcal{D}_L$ such that the distance between any two dictionaries in $\mathcal{D}_L$ is sufficiently large and the hypothesis testing problem is sufficiently hard, i.e., distinct dictionaries result in similar observations. Specifically, for $l, l' \in [L]$, we desire a construction such that

$$\forall l \ne l', \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F \ge 2\sqrt{2\varepsilon} \quad \text{and} \\ D_{KL}\left(f_{\mathbf{D}_l}(\mathbf{Y}) \| f_{\mathbf{D}_{l'}}(\mathbf{Y})\right) \le \alpha_L, \tag{9}$$

where $D_{KL}\left(f_{\mathbf{D}_l}(\mathbf{Y}) \| f_{\mathbf{D}_{l'}}(\mathbf{Y})\right)$ denotes the Kullback-Leibler (KL) divergence between the distributions of observations based on $\mathbf{D}_l \in \mathcal{D}_L$ and $\mathbf{D}_{l'} \in \mathcal{D}_L$, while ε and α_L are non-negative parameters. Roughly, the minimax risk analysis proceeds as follows. Considering $\widehat{\mathbf{D}}(\mathbf{Y})$ to be an estimator that achieves ε^*, and assuming $\varepsilon^* < \varepsilon$ and $\mathbf{D}_l$ generated uniformly at random from $\mathcal{D}_L$, we have $\mathbb{P}(\widehat{l}(\mathbf{Y}) \ne l) = 0$ for the minimum-distance detector $\widehat{l}(\mathbf{Y})$ as long as $\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F < \sqrt{2\varepsilon}$. The goal then is to relate ε^* to $\mathbb{P}(\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F \ge \sqrt{2\varepsilon})$ and $\mathbb{P}(\widehat{l}(\mathbf{Y}) \ne l)$ using Fano's inequality [19]:

$$(1 - \mathbb{P}(\widehat{l}(\mathbf{Y}) \ne l)) \log_2 L - 1 \le I(\mathbf{Y}; l), \tag{10}$$

where $I(\mathbf{Y};l)$ denotes the mutual information (MI) between the observations $\mathbf{Y}$ and the dictionary $\mathbf{D}_l$. Notice that the smaller α_L is in (9), the smaller $I(\mathbf{Y};l)$ will be in (10). Unfortunately, explicitly evaluating $I(\mathbf{Y};l)$ is a challenging task in our setup because of the underlying distributions. Similar to [16], we will instead resort to upper bounding $I(\mathbf{Y};l)$ by assuming access to some side information $\mathbf{T}(\mathbf{X})$ that will make the observations $\mathbf{Y}$ conditionally multivariate Gaussian (recall that $I(\mathbf{Y};l) \leq I(\mathbf{Y};l|\mathbf{T}(\mathbf{X}))$). Our final results will then follow from the fact that any lower bound for ε^* given the side information $\mathbf{T}(\mathbf{X})$ will also be a lower bound for the general case [16].

B. Coefficient distribution

The minimax lower bounds in this paper are derived for various coefficient distributions. First, similar to [16], we consider arbitrary coefficient distributions for which the covariance matrix $\mathbf{\Sigma}_x$ exists. We then specialize our results for sparse coefficient vectors and, under additional assumptions on the reference dictionary $\mathbf{D}_0$, obtain a tighter lower bound for some signal-to-noise ratio (SNR) regimes, where $\text{SNR} = \frac{\mathbf{E_x}(\|\mathbf{Dx}\|_2^2)}{\mathbf{E_n}(\|\mathbf{n}\|_2^2)}$.

1) General coefficients: The coefficient vector $\mathbf{x}$ in this case is assumed to be a zero-mean random vector with covariance $\mathbf{\Sigma}_x$. We also assume access to the side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ to obtain a lower bound on ε^* in this setup.

2) Sparse coefficients: In this case, we assume $\mathbf{x}$ to be an s-sparse vector such that the support of $\mathbf{x}$, denoted by $\text{supp}(\mathbf{x})$, is uniformly distributed over $\mathcal{E} = \{\mathcal{S} \subseteq [p] : |\mathcal{S}| = s\}$:

$$\mathbb{P}(\text{supp}(\mathbf{x}) = \mathcal{S}) = \frac{1}{\binom{p}{s}} \quad \text{for any } \mathcal{S} \in \mathcal{E}. \tag{11}$$

Further, we model the nonzero entries of $\mathbf{x}$, i.e., $\mathbf{x}_\mathcal{S}$, as drawn in an i.i.d. fashion from a distribution with variance σ_a^2:

$$\mathbb{E}_x\{\mathbf{x}_\mathcal{S}\mathbf{x}_\mathcal{S}^T|\mathcal{S}\} = \sigma_a^2\mathbf{I}_s. \tag{12}$$

Notice that an $\mathbf{x}$ under the assumptions of (11) and (12) has

$$\mathbf{\Sigma}_x = \frac{s}{p}\sigma_a^2\mathbf{I}_p. \tag{13}$$

Further, it is easy to see in this case that $\text{SNR} = \frac{s\sigma_a^2}{m\sigma^2}$. Finally, the side information assumed in this sparse coefficients setup will either be $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ or $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$.

III. Lower Bound for General Coefficients

We now provide our main result for the lower bound for the minimax risk of the KS dictionary learning problem for the case of general (i.i.d.) coefficient vectors.

Theorem 1. *Consider a KS dictionary learning problem with N i.i.d observations generated according to model (3) and the true dictionary satisfying (6) for some r and $\mathbf{D}_0$. Suppose $\mathbf{\Sigma}_x$ exists for the zero-mean random coefficient vectors. If there exists an estimator with worst-case MSE $\varepsilon^* \leq \frac{2p(1-t)}{8}\min\{1, \frac{r^2}{4p}\}$, then the minimax risk is lower bounded by*

$$\varepsilon^* \geq \frac{C_1 r^2\sigma^2}{Np\|\mathbf{\Sigma}_x\|_2}\left(c_1(p_1(m_1-1)+p_2(m_2-1))-3\right) \tag{14}$$

for any $0 < c_1 < \frac{t}{8\log 2}$ and $0 < t < 1$, where $C_1 = \frac{(1-t)p}{32r^2}$.

Outline of Proof: The idea of the proof, as discussed in section II-A, is that we construct a set of L distinct KS dictionaries that satisfy:

- $\mathcal{D}_L = \{\mathbf{D}_1, \ldots, \mathbf{D}_L\} \subset \mathcal{X}(\mathbf{D}_0, r)$
- Any two distinct dictionaries in $\mathcal{D}_L$ are separated by a minimum distance in the neighborhood, i.e., for any $l, l' \in [L]$ and some positive $\varepsilon \leq \frac{2p(1-t)}{8}\min\{1, \frac{r^2}{4p}\}$:

$$\|\mathbf{D}_l - \mathbf{D}_l'\|_F \geq 2\sqrt{2\varepsilon}, \text{ for } l \neq l'. \tag{15}$$

Notice that if the true dictionary, $\mathbf{D}_l \in \mathcal{D}_L$, is selected uniformly at random from $\mathcal{D}_L$ in this case then, given side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, the observations $\mathbf{Y}$ follow a multivariate Gaussian distribution and an upper bound on the conditional MI $I(\mathbf{Y};l|\mathbf{T}(\mathbf{X}))$ can be obtained by using an upper bound for KL-divergence of multivariate Gaussian distributions. This bound depends on parameters $\varepsilon, N, m_1, m_2, p_1, p_2, \mathbf{\Sigma}_x, s, r$, and σ^2.

Next, assuming (15) holds for $\mathcal{D}_L$, if there exists an estimator $\widehat{\mathbf{D}}(\mathbf{Y})$ achieving the minimax risk $\varepsilon^* \leq \varepsilon$ and the recovered dictionary $\widehat{\mathbf{D}}(\mathbf{Y})$ satisfies $\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F < \sqrt{2\varepsilon}$, the minimum distance detector $\widehat{l}(\mathbf{Y})$ can recover $\mathbf{D}_l$. Consequently, the probability of error $\mathbb{P}(\widehat{\mathbf{D}}(\mathbf{Y}) \neq \mathbf{D}_l) \leq \mathbb{P}(\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F \geq \sqrt{2\varepsilon})$ can be used to lower bound the conditional MI using Fano's inequality. The obtained lower bound in our case will only be a function of L.

Finally, using the obtained upper and lower bounds for the conditional MI:

$$\eta_2 \leq I(\mathbf{Y};l|\mathbf{T}(\mathbf{X})) \leq \eta_1, \tag{16}$$

a lower bound for the minimax risk ε^* is attained.

A formal proof of Theorem 1 relies on the following lemmas whose proofs appear in the full version of this work [17]. Note that since our construction of $\mathcal{D}_L$ is more complex than the vector case [16, Theorem 1], it requires a different sequence of lemmas, with the exception of Lemma 3, which follows from the vector case.

Lemma 1. *There exists a set of $L = 2^{c_1(mp)-\frac{1}{2}}$ matrices $\mathbf{A}_l \in \mathbb{R}^{m\times p}$, where elements of $\mathbf{A}_l$ take values $\pm\alpha$ for some $\alpha > 0$, such that for $l, l' \in [L]$, $l \neq l'$, any $t > 0$ and $c_1 < \frac{1}{2\log 2}\left(\frac{t}{2\alpha^2 mp}\right)^2$, the following relation is satisfied:*

$$\left|\sum(\mathbf{A}_l \odot \mathbf{A}_{l'})\right| \leq t. \tag{17}$$

Lemma 2. *Considering the generative model in (3), given some $r > 0$ and reference dictionary $\mathbf{D}_0$, there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{c_1((m_1-1)p_1+(m_2-1)p_2)-1}$ such that for any $0 < c_1 < \frac{t^2}{8\log 2}$, any $0 < t < 1$, and any $\varepsilon' > 0$ satisfying*

$$\varepsilon' < \min\left\{r^2, \frac{r^4}{4p}\right\} \tag{18}$$

and any $l, l' \in [L]$, with $l \neq l'$, we have

$$\frac{2p}{r^2}(1-t)\varepsilon' \leq \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \leq \frac{8p}{r^2}\varepsilon'. \tag{19}$$

Furthermore, considering the general coefficient model for $\mathbf{X}$ *and assuming side information* $\mathbf{T}(\mathbf{X}) = \mathbf{X}$*, we have*

$$\forall\, l,\ I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq \frac{4Np\|\mathbf{\Sigma}_x\|_2}{r^2\sigma^2}\varepsilon'. \tag{20}$$

Lemma 3. *Consider the generative model in* (3) *with minimax risk* $\varepsilon^* \leq \varepsilon$ *for some* $\varepsilon > 0$*. Assume there exists a finite set* $\mathcal{D}_L \subseteq \mathcal{D}$ *with* L *dictionaries satisfying*

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \geq 8\varepsilon \tag{21}$$

for $l \neq l'$*. Then for any side information* $\mathbf{T}(\mathbf{X})$*, we have*

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \geq \frac{1}{2}\log_2(L) - 1. \tag{22}$$

Proof of Theorem 1. According to Lemma 2, for any ε' satisfying (18), there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{c_1((m_1-1)p_1+(m_2-1)p_2)-1}$ that satisfies (20) for any $c_1 < \frac{t}{8\log 2}$ and any $0 < t < 1$. According to Lemma 3, if we set $\frac{2p}{r^2}(1-t)\varepsilon' = 8\varepsilon$, (21) is satisfied for $\mathcal{D}_L$ and provided there exists an estimator with worst case MSE satisfying $\varepsilon^* \leq \frac{2p(1-t)}{8}\min\{1, \frac{r^2}{4p}\}$, (22) holds. Combining (20) and (22) we get

$$\frac{1}{2}\log_2(L) - 1 \leq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq \frac{32Np\|\mathbf{\Sigma}_x\|_2}{c_2 r^2\sigma^2}\varepsilon, \tag{23}$$

where $c_2 = \frac{2p}{r^2}(1-t)$. Defining $C_1 = \frac{(1-t)p}{32r^2}$, (23) translates into

$$\varepsilon \geq \frac{C_1 r^2\sigma^2}{Np\|\mathbf{\Sigma}_x\|_2}\left(c_1(p_1(m_1-1) + p_2(m_2-1)) - 3\right). \tag{24}$$

□

IV. Lower Bound for Sparse Coefficients

We now turn our attention to the case of sparse coefficients and obtain lower bounds for the corresponding minimax risk. We first state a corollary of Theorem 1, for $\mathbf{T}(\mathbf{X}) = \mathbf{X}$.

Corollary 1. *Consider a KS dictionary learning problem with* N *i.i.d observations according to model* (3)*. Assuming the true dictionary satisfies* (6) *for some* r *and the reference dictionary* $\mathbf{D}_0$ *satisfies* $\mathsf{RIP}(s, \frac{1}{2})$*, if the random coefficient vector* $\mathbf{x}$ *is selected according to* (11) *and there exists an estimator with worst-case MSE error* $\varepsilon^* \leq \frac{2p(1-t)}{8}\min\{1, \frac{r^2}{4p}\}$*, the minimax risk is lower bounded by*

$$\varepsilon^* \geq \frac{C_1 r^2\sigma^2}{Ns\sigma_a^2}\left(c_1\left(p_1(m_1-1) + p_2(m_2-1)\right) - 3\right) \tag{25}$$

for any $0 < c_1 < \frac{t}{8\log 2}$ *and* $0 < t < 1$*, where* $C_1 = \frac{(1-t)p}{32r^2}$*.*

This result is a direct consequence of Theorem 1, by substituting the covariance matrix of $\mathbf{X}$ given in (13) in (14).

A. Sparse Gaussian coefficients

In this section, we make an additional assumption on the coefficient vector generated according to (11) and assume non-zero elements of $\mathbf{x}$ follow a Gaussian distribution. By additionally assuming the non-zero entries of $\mathbf{x}$ are i.i.d., we can write $\mathbf{x}_S$ as

$$\mathbf{x}_S \sim \mathcal{N}(\mathbf{0}, \sigma_a^2\mathbf{I}_s). \tag{26}$$

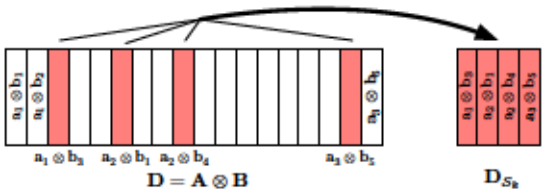

Fig. 1. An illustration of $\mathbf{D}_{l,\mathcal{S}_k}$ with $p_1 = 3$, $p_2 = 6$ and sparsity $s = 4$. Here, $\mathcal{S}_{ka} = \{1, 2, 2, 3\}$, $\mathcal{S}_{kb} = \{3, 1, 4, 5\}$, and $\mathcal{S}_k = \{3, 7, 10, 17\}$.

Therefore, given side information $\mathbf{T}(\mathbf{x}) = \mathrm{supp}(\mathbf{x})$, observations $\mathbf{y}$ follow a multivariate Gaussian distribution. We now provide a theorem for the lower bound attained for this coefficient distribution.

Theorem 2. *Consider a KS dictionary learning problem with* N *i.i.d observations according to model* (3)*. Assuming the true dictionary satisfies* (6) *for some* r *and the reference coordinate dictionaries* $\mathbf{A}_0$ *and* $\mathbf{B}_0$ *satisfy* $\mathsf{RIP}(s, \frac{1}{2})$*, if the random coefficient vector* $\mathbf{x}$ *is selected according to* (11) *and* (26) *and there exists an estimator with worst-case MSE error* $\varepsilon^* \leq \frac{2p(1-t)}{8}\min\{\frac{1}{s}, \frac{r^2}{4p}\}$*, then the minimax risk is lower bounded by*

$$\varepsilon^* \geq \frac{C_2 r^2\sigma^4}{Ns^2\sigma_a^4}\left(c_1(p_1(m_1-1) + p_2(m_2-1)) - 3\right) \tag{27}$$

for any $0 < c_1 < \frac{t}{8\log 2}$ *and* $0 < t < 1$*, where* $C_2 = 1.58 \times 10^{-5}.\frac{p(1-t)}{r^2}$*.*

Outline of Proof: The constructed dictionary class $\mathcal{D}_L$ in Theorem 2 is similar to that in Theorem 1. But the upper bound for the conditional MI, $I(\mathbf{Y}; l|\,\mathrm{supp}(\mathbf{X}))$, differs from that in Theorem 1 as the side information is different.

Given the true dictionary $\mathbf{D}_l$ and support $\mathcal{S}_k$ for the k-th coefficient vector $\mathbf{x}_k$, let $\mathbf{D}_{l,\mathcal{S}_k}$ denote the columns of $\mathbf{D}_l$ corresponding to the non-zeros elements of $\mathbf{x}_k$. In this case, we have

$$\mathbf{y}_k = \mathbf{D}_{l,\mathcal{S}_k}\mathbf{x}_{\mathcal{S}_k} + \mathbf{n}_k,\ k \in [N]. \tag{28}$$

We can write the subdictionary $\mathbf{D}_{l,\mathcal{S}_k}$ in terms of the Khatri-Rao product of two smaller matrices:

$$\mathbf{D}_{l,\mathcal{S}_k} = \mathbf{A}_{l_a,\mathcal{S}_{ka}} * \mathbf{B}_{l_b,\mathcal{S}_{kb}}, \tag{29}$$

where $\mathcal{S}_{ka} = \{i_k\}_{k=1}^s, i_k \in [p_1]$, and $\mathcal{S}_{kb} = \{i'_k\}_{k=1}^s, i'_k \in [p_2]$, are multisets with the following relationship with $\mathcal{S}_k = \{i''_k\}_{k=1}^s, i''_k \in [p]$: $i''_k = (i_k - 1)p_2 + i'_k,\ k \in [s]$. Note that $\mathbf{A}_{l_a,\mathcal{S}_{ka}}$ and $\mathbf{B}_{l_b,\mathcal{S}_{kb}}$ are not submatrices of $\mathbf{A}_{l_a}$ and $\mathbf{B}_{l_b}$, as $\mathcal{S}_{ka}$ and $\mathcal{S}_{kb}$ are multisets. Figure 1 provides a visual illustration of (29). Therefore, the observations follow a multivariate Gaussian distribution with zero mean and covariance matrix:

$$\mathbf{\Sigma}_{k,l} = \sigma_a^2(\mathbf{A}_{l_a,\mathcal{S}_{ka}} * \mathbf{B}_{l_b,\mathcal{S}_{kb}})(\mathbf{A}_{l_a,\mathcal{S}_{ka}} * \mathbf{B}_{l_b,\mathcal{S}_{kb}})^T + \sigma^2\mathbf{I}_s \tag{30}$$

and we need to obtain an upper bound for the conditional MI using (30). We state a variation of Lemma 2 necessary for the proof of Theorem 2. The proof of the lemma is again provided in [17].

Lemma 4. *Considering the generative model in* (3)*, given some* $r > 0$ *and reference dictionary* $\mathbf{D}_0$*, there exists a set*

TABLE I
ORDER-WISE LOWER BOUNDS ON THE MINIMAX RISK FOR VARIOUS COEFFICIENT DISTRIBUTIONS

Distribution \ Dictionary	Unstructured [16]	Kronecker (this paper)
Sparse	$\frac{r^2 p}{N\text{SNR}}$	$\frac{r^2(m_1 p_1 + m_2 p_2)}{Nm\text{SNR}}$
Gaussian Sparse	$\frac{r^2 p}{Nm\text{SNR}^2}$	$\frac{r^2(m_1 p_1 + m_2 p_2)}{Nm^2\text{SNR}^2}$

$\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{c_1((m_1-1)p_1+(m_2-1)p_2)-1}$ such that for any $0 < c_1 < \frac{t^2}{8\log 2}$, any $0 < t < 1$, and any $\varepsilon' > 0$ satisfying

$$0 < \varepsilon' \leq \min\left\{\frac{r^2}{s}, \frac{r^4}{4p}\right\}, \tag{31}$$

and any $l, l' \in [L]$, with $l \neq l'$, we have

$$\frac{2p}{r^2}(1-t)\varepsilon' \leq \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \leq \frac{8p}{r^2}\varepsilon'. \tag{32}$$

Furthermore, assuming the reference coordinate dictionaries $\mathbf{A}_0$ and $\mathbf{B}_0$ satisfy $\text{RIP}(s, \frac{1}{2})$ and the coefficient matrix $\mathbf{X}$ is selected according to (11) and (26), considering side information $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$, we have:

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq 7921\left(\frac{\sigma_a}{\sigma}\right)^4 \frac{Ns^2}{r^2}\varepsilon'. \tag{33}$$

Proof of Theorem 2. According to Lemma 4, for any ε' satisfying (31), there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{c_1((m_1-1)p_1+(m_2-1)p_2)-1}$ that satisfies (33) for any $c_1 < \frac{t}{8\log 2}$ and any $0 < t < 1$. Setting $\frac{2p}{r^2}(1-t)\varepsilon' = 8\varepsilon$, (21) is satisfied for $\mathcal{D}_L$ and, provided there exists an estimator with worst case MSE satisfying $\varepsilon^* \leq \frac{2p(1-t)}{8}\min\{\frac{1}{s}, \frac{r^2}{4p}\}$, (22) holds. Consequently,

$$\frac{1}{2}\log_2(L) - 1 \leq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq \frac{8 \times 7921}{c_2}\left(\frac{\sigma_a}{\sigma}\right)^4 \frac{Ns^2}{r^2}\varepsilon, \tag{34}$$

where $c_2 = \frac{2p}{r^2}(1-t)$. Defining $C_2 = 1.58 \times 10^{-5}.\frac{p(1-t)}{r^2}$, (34) can be written as

$$\varepsilon \geq C_2\left(\frac{\sigma}{\sigma_a}\right)^4 \frac{r^2\left(c_1(p_1(m_1-1)+p_2(m_2-1))-3\right)}{Ns^2}. \tag{35}$$

□

V. DISCUSSION AND CONCLUSION

In this paper we follow an information-theoretic approach to provide lower bounds for the worst-case MSE of KS dictionaries that generate 2-dimensional tensor data. Table I lists the dependence of the known lower bounds on the minimax rates on various parameters of the dictionary learning problem and the SNR$= \frac{s\sigma_a^2}{pn\sigma^2}$. Compared to the results in [16] for the unstructured dictionary learning problem, which are not stated in this form, but can be reduced to this, we are able to decrease the lower bound in all cases by reducing the scaling $\mathcal{O}(pn)$ to $\mathcal{O}(p_1m_1 + p_2m_2)$ for KS dictionaries. This is intuitively pleasing since the minimax lower bound has a linear relationship with the number of degrees of freedom of the KS dictionary, which is $(p_1m_1 + p_2m_2)$, and the square of the neighborhood radius r^2. The results also show that the minimax risk decreases with a larger number of samples N and increased SNR. Notice also that in high SNR regimes, the lower bound in (25) is tighter, while (27) results in a tighter lower bound in low SNR regimes. Our bounds depend on the signal distribution and imply necessary sample complexity scaling $N = \mathcal{O}(r^2(m_1p_1 + m_2p_2)$. Future work includes extending the lower bounds for higher-order tensors and also specifying a learning scheme that achieves these lower bounds.

REFERENCES

[1] 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, 2003.

[2] Z. Zhang and S. Aeron, "Denoising and completion of 3D data via multidimensional dictionary learning," *arXiv preprint arXiv:1512.09227*, 2015.

[3] G. Duan, H. Wang, Z. Liu, J. Deng, and Y.-W. Chen, "K-CPD: Learning of overcomplete dictionaries for tensor sparse coding," in *Proc. IEEE 21st Int. Conf. Pattern Recognition (ICPR)*, 2012, pp. 493–496.

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

[5] 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.

[6] Y. Peng, D. Meng, Z. Xu, C. Gao, Y. Yang, and B. Zhang, "Decomposable nonlocal tensor dictionary learning for multispectral image denoising," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2014, pp. 2949–2956.

[7] S. Soltani, M. E. Kilmer, and P. C. Hansen, "A tensor-based dictionary learning approach to tomographic image reconstruction," *arXiv preprint arXiv:1506.04954*, 2015.

[8] M. Aharon, M. Elad, and A. M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear algebra and its applications*, vol. 416, no. 1, pp. 48–67, 2006.

[9] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli, "Learning sparsely used overcomplete dictionaries via alternating minimization," *arXiv preprint arXiv:1310.7991*, 2013.

[10] A. Agarwal, A. Anandkumar, and P. Netrapalli, "Exact recovery of sparsely used overcomplete dictionaries," *arXiv preprint arXiv:1309.1952.*, 2013.

[11] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Proc. 27th Conf. Learning Theory*, 2014, pp. 779–806.

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

[13] ——, "Local identification of overcomplete dictionaries," *Journal of Machine Learning Research*, vol. 16, pp. 1211–1242, 2015.

[14] R. Gribonval, R. Jenatton, and F. Bach, "Sparse and spurious: dictionary learning with noise and outliers," *arXiv preprint arXiv:1407.5155*, 2014.

[15] A. Jung, Y. C. Eldar, and N. Gortz, "Performance limits of dictionary learning for sparse coding," in *Proc. IEEE 22nd European Signal Process. Conf. (EUSIPCO)*, 2014, pp. 765–769.

[16] A. Jung, Y. C. Eldar, and N. Görtz, "On the minimax risk of dictionary learning," *arXiv preprint arXiv:1507.05498*, 2015.

[17] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds on dictionary learning for tensor data," 2016, preprint.

[18] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[19] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*. Springer, 1997, pp. 423–435.

[20] A. Smilde, R. Bro, and P. Geladi, *Multi-way analysis: Applications in the chemical sciences*. John Wiley & Sons, 2005.

[21] C. F. Van Loan, "The ubiquitous Kronecker product," *Journal of computational and applied mathematics*, vol. 123, no. 1, pp. 85–100, 2000.