Minimax Lower Bounds on Dictionary Learning for Tensor Data

Zahra Shakeri, Waheed U. Bajwa, and Anand D. Sarwate

***Abstract*—This paper provides fundamental limits on the sample complexity of estimating dictionaries for tensor data. The specific focus of this work is on Kth-order tensor data and the case where the underlying dictionary can be expressed in terms of K smaller dictionaries. It is assumed the data are generated by linear combinations of these structured dictionary atoms and observed through white Gaussian noise. This work first provides a general lower bound on the minimax risk of dictionary learning for such tensor data and then adapts the proof techniques for specialized results in the case of sparse and sparse-Gaussian linear combinations. The results suggest the sample complexity of dictionary learning for tensor data can be significantly lower than that for unstructured data: for unstructured data it scales linearly with the product of the dictionary dimensions, whereas for tensor-structured data the bound scales linearly with the sum of the product of the dimensions of the (smaller) component dictionaries. A partial converse is provided for the case of 2nd-order tensor data to show that the bounds in this paper can be tight. This involves developing an algorithm for learning highly-structured dictionaries from noisy tensor data. Finally, numerical experiments highlight the advantages associated with explicitly accounting for tensor data structure during dictionary learning.**

***Index Terms*—Dictionary learning, Kronecker-structured dictionary, minimax bounds, sparse representations, tensor data.**

I. Introduction

Dictionary learning is a technique for finding sparse representations of signals or data and has applications in various tasks such as image denoising and inpainting [3], audio processing [4], and classification [5], [6]. Given input training signals $\{\mathbf{y}_n \in \mathbb{R}^m\}_{n=1}^N$, the goal in dictionary learning is to construct an overcomplete basis, $\mathbf{D} \in \mathbb{R}^{m \times p}$, such that each signal in $\mathbf{Y} = \big[\mathbf{y}_1, \ldots, \mathbf{y}_N\big]$ can be described by a small number of atoms (columns) of $\mathbf{D}$ [7]. This problem can be posed as the following optimization program:

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F \quad \text{subject to } \forall n, \|\mathbf{x}_n\|_0 \leq s, \qquad (1)$$

where $\mathbf{x}_n$ is the coefficient vector associated with $\mathbf{y}_n$, $\|\cdot\|_0$ counts the number of nonzero entries and s is the maximum number of nonzero elements of $\mathbf{x}_n$. Although existing literature has mostly focused on dictionary learning for one-dimensional data [3]–[7], many real-world signals are multidimensional and have a tensor structure: examples include images, videos, and signals produced via magnetic resonance or computed tomography systems. In traditional dictionary learning literature, multidimensional data are converted into one-dimensional data by vectorizing the signals. Such approaches can result in poor sparse representations because they neglect the multidimensional structure of the data [8]. This suggests that it might be useful to keep the original tensor structure of multidimensional data for efficient dictionary learning and reliable subsequent processing.

There have been several algorithms proposed in the literature that can be used to learn structured dictionaries for multidimensional data [8]–[16]. In [9], a Riemannian conjugate gradient method combined with a nonmonotone line search is used to learn structured dictionaries. Other structured dictionary learning works rely on various tensor decomposition methods such as the Tucker decomposition [10], [12]–[14], [17], the CANDECOMP/PARAFAC (CP) decomposition [16], [18], the HOSVD decomposition [11], [19], the t-product tensor factorization [15], and the tensor-SVD [8], [20]. Furthermore learning sums of structured dictionaries can be used to represent tensor data [12], [13].

In this paper, our focus is on theoretical understanding of the fundamental limits of dictionary learning algorithms that explicitly account for the tensor structure of data in terms of *Kronecker structured* (KS) dictionaries. It has been shown that many multidimensional signals can be decomposed into a superposition of separable atoms [21]–[23]. In this case, a sequence of independent transformations on different data dimensions can be carried out using KS matrices. Such matrices have successfully been used for data representation in hyperspectral imaging, video acquisition, and distributed sensing [23].

To the best of our knowledge, none of the prior works on KS dictionary learning [9]–[12] provide an understanding of the sample complexity of KS dictionary learning algorithms. In contrast, we provide lower bounds on the minimax risk of estimating KS dictionaries from tensor data using *any* estimator. These bounds not only provide means of quantifying the performance of existing KS dictionary learning algorithms, but they also hint at the potential benefits of explicitly accounting for tensor structure of data during dictionary learning.

Manuscript received August 29, 2016; revised October 3, 2017 and January 12, 2018; accepted January 14, 2018. This work is supported in part by the National Science Foundation under awards CCF-1525276 and CCF-1453073, and by the Army Research Office under awards W911NF-14-1-0295 and W911NF-17-1-0546. Some of the results reported here were presented at the 2016 IEEE International Symposium on Information Theory (ISIT) [1] and at the 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) [2].

The authors are with the Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, 94 Brett Road, Piscataway, NJ 08854, USA. (Emails: zahra.shakeri@rutgers.edu, waheed.bajwa@rutgers.edu, and anand.sarwate@rutgers.edu)

A. Our Contributions

Our first result is a general lower bound for the mean squared error (MSE) of estimating KS-dictionaries consisting of $K \geq 2$ coordinate dictionaries that sparsely represent Kth-order tensor data. Here, we define the minimax risk to be the worst-case MSE that is attainable by the best dictionary estimator. Our approach uses the standard procedure for lower bounding the minimax risk in nonparametric estimation by connecting it to the maximum probability of error on a carefully constructed multiple hypothesis testing problem [24], [25]: the technical challenge is in constructing an appropriate set of hypotheses. In particular, consider a dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ consisting of the Kronecker product of K coordinate dictionaries $\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}, k \in \{1, \ldots, K\}$, where $m = \prod_{k=1}^{K} m_k$ and $p = \prod_{k=1}^{K} p_k$, that is generated within the radius r neighborhood (taking the Frobenius norm as the distance metric) of a fixed reference dictionary. Our analysis shows that given a sufficiently large r and keeping some other parameters constant, a sample complexity[1] of $N = \Omega(\sum_{k=1}^{K} m_k p_k)$ is necessary for reconstruction of the true dictionary up to a given estimation error. We also provide minimax bounds on the KS dictionary learning problem that hold for the following distributions for the coefficient vectors $\{\mathbf{x}_n\}$:

- $\{\mathbf{x}_n\}$ are independent and identically distributed (i.i.d.) with zero mean and can have any distribution;
- $\{\mathbf{x}_n\}$ are i.i.d. and sparse;
- $\{\mathbf{x}_n\}$ are i.i.d., sparse, and their non-zero elements follow a Gaussian distribution.

Our second contribution is development and analysis of an algorithm to learn dictionaries formed by the Kronecker product of 2 smaller dictionaries, which can be used to represent 2nd-order tensor data. To this end, we show that under certain conditions on the local neighborhood, the proposed algorithm can achieve one of the earlier obtained minimax lower bounds. Based on this, we believe that our lower bound may be tight more generally, but we leave this for future work.

B. Relationship to Previous Work

In terms of relation to prior work, theoretical insights into the problem of dictionary learning have either focused on specific algorithms for non-KS dictionaries [26]–[32] or lower bounds on minimax risk of dictionary learning for one-dimensional data [33], [34]. The former works provide sample complexity results for reliable dictionary estimation based on appropriate minimization criteria. Specifically, given a probabilistic model for sparse coefficients and a finite number of samples, these works find a local minimizer of a nonconvex objective function and show that this minimizer is a dictionary within a given distance of the true dictionary [30]–[32]. In contrast, Jung et al. [33], [34] provide minimax lower bounds for dictionary learning from one-dimensional data under several coefficient vector distributions and discuss a regime where the bounds are tight in the scaling sense for some signal-to-noise (SNR) values. In particular, for a given dictionary $\mathbf{D}$ and sufficiently large neighborhood radius r, they show that $N = \Omega(mp)$ samples are required for reliable recovery of the dictionary up to a prescribed MSE within its local neighborhood. However, in the case of tensor data, their approach does not exploit the structure in the data, whereas our goal is to show how structure can potentially yield a lower sample complexity in the dictionary learning problem.

To provide lower bounds on the minimax risk of KS dictionary learning, we adopt the same general approach that Jung et al. [33], [34] use for the vector case. They use the standard approach of connecting the estimation problem to a multiple-hypothesis testing problem and invoking Fano's inequality [25]. We construct a family of KS dictionaries which induce similar observation distributions but have a minimum separation from each other. By explicitly taking into account the Kronecker structure of the dictionaries, we show that the sample complexity satisfies a lower bound of $\Omega(\sum_{k=1}^{K} m_k p_k)$ compared to the $\Omega(mp)$ bound from vectorizing the data [34]. Although our general approach is similar to that in [34], there are fundamental differences in the construction of the KS dictionary class and analysis of the minimax risk. This generalizes our preliminary work [1] from 2nd-order to Kth-order and provides a comprehensive analysis of the KS dictionary class construction and minimax lower bounds.

Our results essentially show that the sample complexity depends linearly on the degrees of freedom of a Kronecker structured dictionary, which is $\sum_{k=1}^{K} m_k p_k$, and non-linearly on the SNR and tensor order K. These lower bounds also depend on the radius of the local neighborhood around a fixed reference dictionary. Our results hold even when some of the coordinate dictionaries are not overcomplete[2]. Like the previous work [34], our analysis is local and our lower bounds depend on the distribution of multidimensional data.

We next introduce a KS dictionary learning algorithm for 2nd-order tensor data and show that in this case, one of the provided minimax lower bounds is achievable under certain conditions. We also conduct numerical experiments that demonstrate the empirical performance of the algorithm relative to the MSE upper bound and in comparison to the performance of a non-KS dictionary learning algorithm [34].

C. Notational Convention and Preliminaries

Underlined bold upper-case, bold upper-case and lower-case letters are used to denote real-valued tensors, matrices and vectors, respectively. Lower-case letters denote scalars. The k-th column of $\mathbf{X}$ is denoted by $\mathbf{x}_k$ and its ij-th element is denoted by x_{ij}. Sometimes we use matrices indexed by multiple letters, such as $\mathbf{X}_{(a,b,c)}$, in which case its j-th column is denoted by $\mathbf{x}_{(a,b,c),j}$. The function $\mathrm{supp}(.)$ denotes the locations of the nonzero entries of $\mathbf{X}$. Let $\mathbf{X}_{\mathcal{I}}$ be the matrix consisting of columns of $\mathbf{X}$ with indices $\mathcal{I}$, $\mathbf{X}^{\mathcal{T}}$ be the matrix consisting of rows of $\mathbf{X}$ with indices $\mathcal{T}$ and $\mathbf{I}_d$ be

[1] We use $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$ if for sufficiently large $n \in \mathbb{N}$, $f(n) < C_1 g(n)$ and $f(n) > C_2 g(n)$, respectively, for some positive constants C_1 and C_2.

[2] Note that all coordinate dictionaries cannot be undercomplete, otherwise $\mathbf{D}$ won't be overcomplete.

the $d \times d$ identity matrix. For a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$, its $(i_1, \ldots, i_K)$-th element is denoted as $\underline{x}_{i_1 \ldots i_K}$. Norms are given by subscripts, so $\|\mathbf{u}\|_0$ and $\|\mathbf{u}\|_2$ are the ℓ_0 and ℓ_2 norms of $\mathbf{u}$, respectively, and $\|\mathbf{X}\|_2$ and $\|\mathbf{X}\|_F$ are the spectral and Frobenius norms of $\mathbf{X}$, respectively. We use $\mathrm{vec}(\mathbf{X})$ to denote the vectorized version of matrix $\mathbf{X}$, which is a column vector obtained by stacking the columns of $\mathbf{X}$ on top of one another. We write $[K]$ for $\{1, \ldots, K\}$. For matrices $\mathbf{X}$ and $\mathbf{Y}$, we define their distance in terms of the Frobenius norm:

$$d(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|_F.$$

We define the outer product of two vectors of the same dimension, $\mathbf{u}$ and $\mathbf{v}$, as $\mathbf{u} \odot \mathbf{v} = \mathbf{u}\mathbf{v}^\top$ and the inner product between matrices of the same size, $\mathbf{X}$ and $\mathbf{Y}$, as $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{Tr}(\mathbf{X}^\top \mathbf{Y})$. Furthermore, $P_{\mathcal{B}_1}(\mathbf{u})$ denotes the projection of $\mathbf{u}$ on the closed unit ball, i.e.,

$$P_{\mathcal{B}_1}(\mathbf{u}) = \begin{cases} \mathbf{u}, & \text{if } \|\mathbf{u}\|_2 \leq 1, \\ \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, & \text{otherwise.} \end{cases} \tag{2}$$

We now define some important matrix products. We write $\mathbf{X} \otimes \mathbf{Y}$ for the *Kronecker product* of two matrices $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\mathbf{Y} \in \mathbb{R}^{p \times q}$, defined as

$$\mathbf{X} \otimes \mathbf{Y} = \begin{bmatrix} x_{11}\mathbf{Y} & x_{12}\mathbf{Y} & \ldots & x_{1n}\mathbf{Y} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}\mathbf{Y} & x_{m2}\mathbf{Y} & \ldots & x_{mn}\mathbf{Y} \end{bmatrix}, \tag{3}$$

where the result is an $mp \times nq$ matrix and we have $\|\mathbf{X} \otimes \mathbf{Y}\|_F = \|\mathbf{X}\|_F \|\mathbf{Y}\|_F$ [35]. Given matrices $\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_1$, and $\mathbf{Y}_2$, where products $\mathbf{X}_1\mathbf{Y}_1$ and $\mathbf{X}_2\mathbf{Y}_2$ can be formed, we have [36]

$$(\mathbf{X}_1 \otimes \mathbf{X}_2)(\mathbf{Y}_1 \otimes \mathbf{Y}_2) = (\mathbf{X}_1\mathbf{Y}_1) \otimes (\mathbf{X}_2\mathbf{Y}_2). \tag{4}$$

Given $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\mathbf{Y} \in \mathbb{R}^{p \times n}$, we write $\mathbf{X} * \mathbf{Y}$ for their $mp \times n$ *Khatri-Rao product* [36], defined by

$$\mathbf{X} * \mathbf{Y} = \begin{bmatrix} \mathbf{x}_1 \otimes \mathbf{y}_1 & \mathbf{x}_2 \otimes \mathbf{y}_2 & \ldots & \mathbf{x}_n \otimes \mathbf{y}_n \end{bmatrix}. \tag{5}$$

This is essentially the column-wise Kronecker product of matrices $\mathbf{X}$ and $\mathbf{Y}$. We also use $\bigotimes_{k \in K} \mathbf{X}_k = \mathbf{X}_1 \otimes \cdots \otimes \mathbf{X}_K$ and $\divideontimes_{k \in K} \mathbf{X}_k = \mathbf{X}_1 * \cdots * \mathbf{X}_K$.

Next, we review essential properties of Kth-order tensors and the relation between tensors and the Kronecker product of matrices using the *Tucker decomposition* of tensors.

1) A Brief Review of Tensors: A tensor is a multidimensional array where the order of the tensor is defined as the number of components in the array. A tensor $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_K}$ of order K can be expressed as a matrix by reordering its elements to form a matrix. This reordering is called unfolding: the mode-k unfolding matrix of a tensor is a $p_k \times \prod_{i \neq k} p_i$ matrix, which we denote by $\mathbf{X}_{(k)}$. Each column of $\mathbf{X}_{(k)}$ consists of the vector formed by fixing all indices of $\underline{\mathbf{X}}$ except the one in the kth-order. For example, for a 2nd-order tensor $\underline{\mathbf{X}}$, the mode-1 and mode-2 unfolding matrices are $\underline{\mathbf{X}}$ and $\underline{\mathbf{X}}^\top$, respectively. The k-rank of a tensor $\underline{\mathbf{X}}$ is defined by $\mathrm{rank}(\mathbf{X}_{(k)})$; trivially, $\mathrm{rank}(\mathbf{X}_{(k)}) \leq p_k$.

The mode-k matrix product of the tensor $\underline{\mathbf{X}}$ and a matrix $\mathbf{A} \in \mathbb{R}^{m_k \times p_k}$, denoted by $\underline{\mathbf{X}} \times_k \mathbf{A}$, is a tensor of size $p_1 \times \ldots p_{k-1} \times m_k \times p_{k+1} \cdots \times p_K$ whose elements are

$$(\underline{\mathbf{X}} \times_k \mathbf{A})_{i_1 \ldots i_{k-1} j i_{k+1} \ldots i_K} = \sum_{i_k=1}^{p_k} \underline{x}_{i_1 \ldots i_{k-1} i_k i_{k+1} \ldots i_K} a_{j i_k}. \tag{6}$$

The mode-k matrix product of $\underline{\mathbf{X}}$ and $\mathbf{A}$ and the matrix multiplication of $\mathbf{X}_{(k)}$ and $\mathbf{A}$ are related [37]:

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_k \mathbf{A} \Leftrightarrow \mathbf{Y}_{(k)} = \mathbf{A}\mathbf{X}_{(k)}. \tag{7}$$

2) Tucker Decomposition for Tensors: The Tucker decomposition is a powerful tool that decomposes a tensor into a *core tensor* multiplied by a matrix along each mode [17], [37]. We take advantage of the Tucker model since we can relate the Tucker decomposition to the Kronecker representation of tensors [38]. For the tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$ of order K, if $\mathrm{rank}(\mathbf{Y}_{(k)}) \leq p_k$ holds for all $k \in [K]$ then, according to the Tucker model, $\underline{\mathbf{Y}}$ can be decomposed into:

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times_3 \cdots \times_K \mathbf{D}_K, \tag{8}$$

where $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_K}$ denotes the core tensor and $\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}$ are factor matrices. Here, (8) can be interpreted as a form of higher order principal component analysis (PCA):

$$\underline{\mathbf{Y}} = \sum_{i_1 \in [p_1]} \cdots \sum_{i_K \in [p_K]} \underline{x}_{i_1 \ldots i_K} \mathbf{d}_{1, i_1} \odot \cdots \odot \mathbf{d}_{K, i_K}, \tag{9}$$

where the $\mathbf{D}_k$'s can be interpreted as the principal components in mode-k. The following is implied by (8) [37]:

$$\mathbf{Y}_{(k)} = \mathbf{D}_k \mathbf{X}_{(k)} (\mathbf{D}_K \otimes \cdots \otimes \mathbf{D}_{k+1} \otimes \mathbf{D}_{k-1} \otimes \cdots \otimes \mathbf{D}_1)^\top. \tag{10}$$

Since the Kronecker product satisfies $\mathrm{vec}(\mathbf{B}\mathbf{X}\mathbf{A}^\top) = (\mathbf{A} \otimes \mathbf{B})\,\mathrm{vec}(\mathbf{X})$, (8) is equivalent to

$$\mathrm{vec}(\underline{\mathbf{Y}}) = \left(\mathbf{D}_K \otimes \mathbf{D}_{K-1} \otimes \cdots \otimes \mathbf{D}_1\right) \mathrm{vec}(\underline{\mathbf{X}}), \tag{11}$$

where $\mathrm{vec}(\underline{\mathbf{Y}}) \triangleq \mathrm{vec}(\mathbf{Y}_{(1)})$ and $\mathrm{vec}(\underline{\mathbf{X}}) \triangleq \mathrm{vec}(\mathbf{X}_{(1)})$ [37]–[39].

The rest of the paper is organized as follows. We formulate the KS dictionary learning problem and describe the procedure for obtaining minimax risk lower bounds in Section II. Next, we provide a lower bound for general coefficient distribution in Section III and in Section IV, we present lower bounds for sparse and sparse Gaussian coefficient vectors. We propose a KS dictionary learning algorithm for 2nd-order tensor data and analyze its corresponding MSE and empirical performance in Section V. In Section VI, we discuss and interpret the results. Finally, in Section VII, we conclude the paper. In order to keep the main exposition simple, proofs of most of the lemmas and theorems are relegated to the appendix.

II. Problem Formulation

In the conventional dictionary learning model, it is assumed that the observations $\mathbf{y}_n \in \mathbb{R}^m$ are generated via a fixed dictionary as

$$\mathbf{y}_n = \mathbf{D}\mathbf{x}_n + \boldsymbol{\eta}_n, \tag{12}$$

in which the dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ is an overcomplete basis ($m < p$) with unit-norm columns[3] and rank m, $\mathbf{x}_n \in \mathbb{R}^p$ is the coefficient vector, and $\boldsymbol{\eta}_n \in \mathbb{R}^m$ denotes observation noise.

Our focus in this work is on multidimensional signals. We assume the observations are Kth-order tensors $\underline{\mathbf{Y}}_n \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$. According to the Tucker model, given *coordinate dictionaries* $\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}$, a *coefficient tensor* $\underline{\mathbf{X}}_n \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_K}$, and a *noise tensor* $\underline{\mathbf{N}}_n$, we can write $\mathbf{y}_n \triangleq \mathrm{vec}(\underline{\mathbf{Y}}_n)$ using (11) as[4]

$$\mathbf{y}_n = \left(\bigotimes_{k \in [K]} \mathbf{D}_k \right) \mathbf{x}_n + \boldsymbol{\eta}_n, \tag{13}$$

where $\mathbf{x}_n \triangleq \mathrm{vec}(\underline{\mathbf{X}}_n)$ and $\boldsymbol{\eta}_n \triangleq \mathrm{vec}(\underline{\mathbf{N}}_n)$. Let

$$m = \prod_{k \in [K]} m_k \quad \text{and} \quad p = \prod_{k \in [K]} p_k. \tag{14}$$

Concatenating N i.i.d. noisy observations $\{\mathbf{y}_n\}_{n=1}^N$, which are realizations according to the model (13), into $\mathbf{Y} \in \mathbb{R}^{m \times N}$, we obtain

$$\mathbf{Y} = \mathbf{D}\mathbf{X} + \mathbf{N}, \tag{15}$$

where $\mathbf{D} \triangleq \bigotimes_{k \in [K]} \mathbf{D}_k$ is the unknown KS dictionary, $\mathbf{X} \in \mathbb{R}^{p \times N}$ is a coefficient matrix consisting of i.i.d. random coefficient vectors with known distribution that has zero-mean and covariance matrix $\boldsymbol{\Sigma}_x$, and $\mathbf{N} \in \mathbb{R}^{m \times N}$ is assumed to be additive white Gaussian noise (AWGN) with zero mean and variance σ^2.

Our main goal in this paper is to derive necessary conditions under which the KS dictionary $\mathbf{D}$ can possibly be learned from the noisy observations given in (15). 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

$$\mathbf{D}_0 = \bigotimes_{k \in [K]} \mathbf{D}_{(0,k)}, \tag{16}$$

and $\mathbf{D}_0 \in \mathcal{D}$, where

$$\mathcal{D} \triangleq \Bigg\{ \mathbf{D}' \in \mathbb{R}^{m \times p} : \ \mathbf{D}' = \bigotimes_{k \in [K]} \mathbf{D}'_k, \mathbf{D}'_k \in \mathbb{R}^{m_k \times p_k}, \\ \|\mathbf{d}'_{k,j}\|_2 = 1 \ \forall k \in [K], j \in [p_k] \Bigg\}. \tag{17}$$

We assume the true generating KS dictionary $\mathbf{D}$ belongs to a neighborhood around $\mathbf{D}_0$:

$$\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r) \triangleq \{\mathbf{D}' \in \mathcal{D} : \|\mathbf{D}' - \mathbf{D}_0\|_F < r\} \tag{18}$$

for some fixed radius r.[5] Note that $\mathbf{D}_0$ appears in the analysis as an artifact of our proof technique to construct the dictionary class. In particular, if r is sufficiently large, then $\mathcal{X}(\mathbf{D}_0, r) \approx \mathcal{D}$ and effectively $\mathbf{D} \in \mathcal{D}$.

[3]The unit-norm condition on columns of $\mathbf{D}$ is required to avoid solutions with arbitrary large norms for dictionary columns and small values for $\mathbf{X}$.

[4]We have reindexed $\mathbf{D}_k$'s in (11) for ease of notation.

[5]Note that our results hold with the unit-norm condition enforced only on $\mathbf{D}$ itself, and not on the subdictionaries $\mathbf{D}_k$. Nevertheless, we include this condition in the dictionary class for the sake of completeness as it also ensures uniqueness of the subdictionaries (factors of a K-fold Kronecker product can exchange scalars γ_k freely without changing the product as long as $\prod_{k \in [K]} \gamma_k = 1$).

A. Minimax Risk

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,

$$\varepsilon^* = \inf_{\widehat{\mathbf{D}}} \sup_{\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r)} \mathbb{E}_{\mathbf{Y}} \left\{ \left\| \widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D} \right\|_F^2 \right\}, \tag{19}$$

where $\widehat{\mathbf{D}}(\mathbf{Y})$ can be estimated using any KS dictionary learning algorithm. In order to lower bound this minimax risk ε^*, we employ a standard reduction to the multiple hypothesis testing used in the literature on nonparametric estimation [24], [25]. This approach is equivalent to generating a KS dictionary $\mathbf{D}_l$ uniformly at random from a carefully constructed class $\mathcal{D}_L = \{\mathbf{D}_1, \ldots, \mathbf{D}_L\} \subseteq \mathcal{X}(\mathbf{D}_0, r), L \geq 2$, for a given ($\mathbf{D}_0$, r). 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 large but the hypothesis testing problem is hard; that is, two distinct dictionaries $\mathbf{D}_l$ and $\mathbf{D}_{l'}$ should produce similar observations. Specifically, for $l, l' \in [L]$, and given error $\varepsilon \geq \varepsilon^*$, we desire a construction such that

$$\forall l \neq l', \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F \geq 2\sqrt{\gamma \varepsilon} \quad \text{and} \\ D_{KL}\left(f_{\mathbf{D}_l}(\mathbf{Y}) || f_{\mathbf{D}_{l'}}(\mathbf{Y}) \right) \leq \alpha_L, \tag{20}$$

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 γ, α_L, and ε are non-negative parameters. Observations $\mathbf{Y} = \mathbf{D}_l\mathbf{X} + \mathbf{N}$ in this setting can be interpreted as channel outputs that are used to estimate the input $\mathbf{D}_l$ using an arbitrary KS dictionary algorithm that is assumed to achieve the error ε. Our goal is to detect the correct generating KS dictionary index l. For this purpose, a minimum distance detector is used:

$$\widehat{l} = \min_{l' \in [L]} \left\| \widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_{l'} \right\|_F. \tag{21}$$

Then, we have $\mathbb{P}(\widehat{l}(\mathbf{Y}) \neq l) = 0$ for the minimum-distance detector $\widehat{l}(\mathbf{Y})$ as long as $\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F < \sqrt{\gamma\varepsilon}$. The goal then is to relate ε to $\mathbb{P}(\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F \geq \sqrt{\gamma\varepsilon})$ and $\mathbb{P}(\widehat{l}(\mathbf{Y}) \neq l)$ using Fano's inequality [25]:

$$(1 - \mathbb{P}(\widehat{l}(\mathbf{Y}) \neq l)) \log_2 L - 1 \leq I(\mathbf{Y}; l), \tag{22}$$

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 (20), the smaller $I(\mathbf{Y}; l)$ will be in (22). Unfortunately, explicitly evaluating $I(\mathbf{Y}; l)$ is a challenging task in our setup because the underlying distributions are mixture of distributions. Similar to [34], we will instead resort to upper bounding $I(\mathbf{Y}; l)$ by conditioning it on some side information $\mathbf{T}(\mathbf{X})$ that will make the observations $\mathbf{Y}$ conditionally multivariate Gaussian (in particular, from [34, Lemma

A.1], it follows that $I(\mathbf{Y};l) \leq I(\mathbf{Y};l|\mathbf{T}(\mathbf{X}))$.[6] We will in particular focus on two types of side information: $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ and $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$. 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 $\{m_k\}_{k=1}^K$ and $\{p_k\}_{k=1}^K$ of the true KS dictionary, neighborhood radius r, and coefficient covariance $\mathbf{\Sigma}_x$, but also on the structure of the constructed class $\mathcal{D}_L$ [24]. Note that our approach is applicable to the global KS dictionary learning problem, since the minimax lower bounds that are obtained for any $\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r)$ are also trivially lower bounds for $\mathbf{D} \in \mathcal{D}$.

After providing minimax lower bounds for the KS dictionary learning problem, we develop and analyze a simple KS dictionary learning algorithm for $K = 2$ order tensor data. Our analysis shows that one of our provided lower bounds is achievable, suggesting that they may be tight.

B. Coefficient Distribution

By making different assumptions on coefficient distributions, we can specialize our lower bounds to specific cases. To facilitate comparisons with prior work, we adopt somewhat similar coefficient distributions as in the unstructured case [34]. First, we consider any coefficient distribution and only assume that the coefficient covariance matrix exists. We then specialize our analysis to sparse coefficient vectors and, by adding additional conditions on the reference dictionary $\mathbf{D}_0$, we obtain a tighter lower bound for the minimax risk for some SNR regimes.

1) General Coefficients: First, we consider the general case, where $\mathbf{x}$ is a zero-mean random coefficient vector with covariance matrix $\mathbf{\Sigma}_x = \mathbb{E}_{\mathbf{x}}\left\{\mathbf{x}\mathbf{x}^\top\right\}$. We make no additional assumption on the distribution of $\mathbf{x}$. We condition on side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ to obtain a lower bound on the minimax risk in the case of general coefficients.

2) Sparse Coefficients: In the case where the coefficient vector is sparse, we show that additional assumptions on the non-zero entries yield a lower bound on the minimax risk conditioned on side information $\text{supp}(\mathbf{x})$, which denotes the support of $\mathbf{x}$ (the set containing indices of the locations of the nonzero entries of $\mathbf{x}$). We study two cases for the distribution of $\text{supp}(\mathbf{x})$:

- **Random Sparsity.** In this case, the random support of $\mathbf{x}$ is distributed uniformly over $\mathcal{E}_1 = \{\mathcal{S} \subseteq [p] : |\mathcal{S}| = s\}$:

$$\mathbb{P}(\text{supp}(\mathbf{x}) = \mathcal{S}) = \frac{1}{\binom{p}{s}}, \quad \text{for any } \mathcal{S} \in \mathcal{E}_1. \tag{23}$$

- **Separable Sparsity.** In this case we sample s_k elements uniformly at random from $[p_k]$, for all $k \in [K]$. The random support of $\mathbf{x}$ is $\mathcal{E}_2 = \{\mathcal{S} \subseteq [p] : |\mathcal{S}| = s\}$, where $\mathcal{S}$ is related to $\{\mathcal{S}_1 \times \cdots \times \mathcal{S}_K : \mathcal{S}_k \subseteq [p_k], |\mathcal{S}_k| = s_k, k \in [K]\}$ via lexicographic indexing. The number of non-zero elements in $\mathbf{x}$ in this case is $s = \prod_{k\in[K]} s_k$. The probability of sampling K subsets $\{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ is

$$\mathbb{P}(\text{supp}(\mathbf{x}) = \mathcal{S}) = \frac{1}{\prod_{k\in[K]} \binom{p_k}{s_k}}, \quad \text{for any } \mathcal{S} \in \mathcal{E}_2. \tag{24}$$

In other words, separable sparsity requires non-zero coefficients to be grouped in blocks. This model arises in the case of processing of images and video sequences [38].

Remark 1. If $\underline{\mathbf{X}}$ follows the separable sparsity model with sparsity $(s_1, \ldots, s_K)$, then the columns of the mode-k matrix $\mathbf{Y}_{(k)}$ of $\underline{\mathbf{Y}}$ have s_k-sparse representations with respect to $\mathbf{D}_k$, for $k \in [K]$ [38].

For a signal $\mathbf{x}$ with sparsity pattern $\text{supp}(\mathbf{x})$, we model the non-zero entries of $\mathbf{x}$, i.e., $\mathbf{x}_{\mathcal{S}}$, as drawn independently and identically from a probability distribution with known variance σ_a^2:

$$\mathbb{E}_x\{\mathbf{x}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^T|\mathcal{S}\} = \sigma_a^2\mathbf{I}_s. \tag{25}$$

Any $\mathbf{x}$ with sparsity model (23) or (24) and nonzero entries satisfying (25) has covariance matrix

$$\mathbf{\Sigma}_x = \frac{s}{p}\sigma_a^2\mathbf{I}_p. \tag{26}$$

III. Lower Bound for General Distribution

We now provide our main result for the lower bound for minimax risk of the KS dictionary learning problem for the case of general coefficient distributions.

Theorem 1. *Consider a KS dictionary learning problem with N i.i.d. observations generated according to model* (13)*. Suppose the true dictionary satisfies* (18) *for some r and fixed reference dictionary $\mathbf{D}_0$ satisfying* (16)*. Then for any coefficient distribution with mean zero and covariance $\mathbf{\Sigma}_x$, we have the following lower bound on ε^*:*

$$\varepsilon^* \geq \frac{t}{4}\min\left\{p, \frac{r^2}{2K}, \frac{\sigma^2}{4NK\|\mathbf{\Sigma}_x\|_2}\left(c_1\left(\sum_{k\in[K]}(m_k-1)p_k\right) - \frac{K}{2}\log_2 2K - 2\right)\right\}, \tag{27}$$

for any $0 < t < 1$ and any $0 < c_1 < \dfrac{1-t}{8\log 2}$.

The implications of Theorem 1 are examined in Section VI.

Outline of Proof: The idea of the proof is that we construct a set of L distinct KS dictionaries, $\mathcal{D}_L = \{\mathbf{D}_1, \ldots, \mathbf{D}_L\} \subset \mathcal{X}(\mathbf{D}_0, r)$, such that any two distinct dictionaries are separated by a minimum distance. That is for any pair $l, l' \in [L]$ and any positive $\varepsilon < \dfrac{tp}{4}\min\left\{r^2, \dfrac{r^4}{2Kp}\right\}$:

$$\|\mathbf{D}_l - \mathbf{D}_l'\|_F \geq 2\sqrt{2\varepsilon}, \text{ for } l \neq l'. \tag{28}$$

In this case, if a dictionary $\mathbf{D}_l \in \mathcal{D}_L$ is selected uniformly at random from $\mathcal{D}_L$, then conditioned on side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, the observations under this dictionary follow a multivariate Gaussian distribution. We can therefore upper bound the conditional MI by approximating the upper bound for KL-divergence of multivariate

[6] Instead of upper bounding $I(\mathbf{Y};l|\mathbf{T}(\mathbf{X}))$, similar results can be derived by using Fano's inequality for the conditional probability of error, $\mathbb{P}(\widehat{l}(\mathbf{Y}) \neq l|\mathbf{T}(\mathbf{X}))$ [40, Theorem 2].

Gaussian distributions. This bound depends on parameters $\varepsilon, N, \{m_k\}_{k=1}^K, \{p_k\}_{k=1}^K, \mathbf{\Sigma}_x, s, r, K$, and σ^2.

Assuming (28) holds for $\mathcal{D}_L$, if there exists an estimator 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 can recover $\mathbf{D}_l$. Then, using the Markov inequality and since ε^* is bounded, 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 upper bounded by $\frac{1}{2}$. Further, according to (22), the lower bound for the conditional MI can be obtained using Fano's inequality [34]. The lower bound is a function of L only. Finally, using the obtained bounds for the conditional MI, we derive a lower bound for the minimax risk ε^*.

Remark 2. We use the constraint in (28) in Theorem 1 for simplicity: the number $2\sqrt{2}$ can be replaced with any arbitrary $\gamma > 0$.

The complete technical proof of Theorem 1 relies on the following lemmas, which are formally proved in the appendix. Although the similarity of our model to that of Jung et al. [34] suggests that our proof should be a simple extension of their proof of Theorem 1, the construction for KS dictionaries is more complex and its analysis requires a different approach. One exception is Lemma 3 [34, Lemma 8], which connects a lower bound on the Frobenius norms of pairwise differences in the construction to a lower bound on the conditional MI used in Fano's inequality [25].

Lemma 1. *Let $\alpha > 0$ and $\beta > 0$. Let $\{\mathbf{A}_l \in \mathbb{R}^{m\times p} : l \in [L]\}$ be a set of L matrices where each $\mathbf{A}_l$ contains $m \times p$ independent and identically distributed random variables taking values $\pm\alpha$ uniformly. Then we have the following inequality:*

$$\mathbb{P}\left(\exists(l,l') \in [L]\times[L], l \neq l' : |\langle \mathbf{A}_l, \mathbf{A}_{l'}\rangle| \geq \beta\right) \leq 2L^2 \exp\left(-\frac{\beta^2}{4\alpha^4 mp}\right). \tag{29}$$

Lemma 2. *Consider the generative model in* (13)*. Fix $r > 0$ and a reference dictionary $\mathbf{D}_0$ satisfying* (16)*. Then there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{\lfloor c_1(\sum_{k\in[K]}(m_k-1)p_k) - \frac{K}{2}\log_2(2K)\rfloor}$ such that for any $0 < t < 1$, any $0 < c_1 < \frac{t^2}{8\log 2}$, any $\varepsilon' > 0$ satisfying*

$$\varepsilon' < r^2 \min\left\{1, \frac{r^2}{2Kp}\right\}, \tag{30}$$

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

$$\frac{2p}{r^2}(1-t)\varepsilon' \leq \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \leq \frac{4Kp}{r^2}\varepsilon'. \tag{31}$$

Furthermore, if $\mathbf{X}$ is drawn from a distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_x$ and conditioning on side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, we have

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq \frac{2NKp\|\mathbf{\Sigma}_x\|_2}{r^2\sigma^2}\varepsilon'. \tag{32}$$

Lemma 3 (Lemma 8 [34])**.** *Consider the generative model in* (13) *and suppose the minimax risk ε^* satisfies $\varepsilon^* \leq \varepsilon$ for some $\varepsilon > 0$. If there exists a finite set $\mathcal{D}_L \subseteq \mathcal{D}$ with L dictionaries satisfying*

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \geq 8\varepsilon \tag{33}$$

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

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \geq \frac{1}{2}\log_2(L) - 1. \tag{34}$$

Proof of Lemma 3: The proof of Lemma 3 is identical to the proof of Lemma 8 in Jung et al. [34]. ∎

Proof of Theorem 1: According to Lemma 2, for any ε' satisfying (30), there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{\lfloor c_1(\sum_{k\in[K]}(m_k-1)p_k) - \frac{K}{2}\log_2(2K)\rfloor}$ that satisfies (32) for any $0 < t' < 1$ and any $c_1 < \dfrac{t'}{8\log 2}$. Let $t = 1 - t'$. If there exists an estimator with worst-case MSE satisfying $\varepsilon^* \leq \dfrac{2tp}{8}\min\left\{1, \dfrac{r^2}{2Kp}\right\}$ then, according to Lemma 3, if we set $\frac{2tp}{r^2}\varepsilon' = 8\varepsilon^*$, (33) is satisfied for $\mathcal{D}_L$ and (34) holds. Combining (32) and (34) we get

$$\frac{1}{2}\log_2(L) - 1 \leq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq \frac{16NKp\|\mathbf{\Sigma}_x\|_2}{c_2 r^2\sigma^2}\varepsilon^*, \tag{35}$$

where $c_2 = \dfrac{2tp}{r^2}$. We can write (35) as

$$\varepsilon^* \geq \frac{t\sigma^2}{16NK\|\mathbf{\Sigma}_x\|_2}\left(c_1\left(\sum_{k\in[K]}(m_k-1)p_k\right) - \frac{K}{2}\log_2 2K - 2\right). \tag{36}$$

∎

IV. Lower Bound for Sparse Distributions

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 sparse coefficients, corresponding to $\mathbf{T}(\mathbf{X}) = \mathbf{X}$.

Corollary 1. *Consider a KS dictionary learning problem with N i.i.d. observations generated according to model* (13)*. Suppose the true dictionary satisfies* (18) *for some r and fixed reference dictionary $\mathbf{D}_0$ satisfying* (16)*. If the random coefficient vector $\mathbf{x}$ is selected according to* (23) *or* (24)*, we have the following lower bound on ε^*:*

$$\varepsilon^* \geq \frac{t}{4}\min\left\{p, \frac{r^2}{2K}, \frac{\sigma^2 p}{4NKs\sigma_a^2}\left(c_1\left(\sum_{k\in[K]}(m_k-1)p_k\right) - \frac{K}{2}\log_2 2K - 2\right)\right\}, \tag{37}$$

for any $0 < t < 1$ and any $0 < c_1 < \dfrac{1-t}{8\log 2}$.

This result is a direct consequence of Theorem 1, obtained by substituting the covariance matrix of sparse coefficients given in (26) into (27).

A. Sparse Gaussian Coefficients

In this section, we make an additional assumption on the coefficient vectors generated according to (23) and assume

non-zero elements of the vectors follow a Gaussian distribution. By additionally assuming the non-zero entries of $\mathbf{x}$ are i.i.d. Gaussian distributed, we can write $\mathbf{x}_{\mathcal{S}}$ as

$$\mathbf{x}_{\mathcal{S}} \sim \mathcal{N}(\mathbf{0}, \sigma_a^2 \mathbf{I}_s). \tag{38}$$

As a result, conditioned on side information $\mathbf{T}(\mathbf{x}_n) = \text{supp}(\mathbf{x}_n)$, observations $\mathbf{y}_n$ follow a multivariate Gaussian distribution. Part of our forthcoming analysis relies on the notion of the *restricted isometry property* (RIP) for a matrix.

*Restricted Isometry Property (*RIP*) [41]:* A matrix $\widetilde{\mathbf{D}}$ with unit ℓ_2-norm columns satisfies the RIP of order s with constant δ_s if

$$(1-\delta_s)\|\mathbf{x}\|_2^2 \leq \|\widetilde{\mathbf{D}}\mathbf{x}\|_2^2 \leq (1+\delta_s)\|\mathbf{x}\|_2^2, \tag{39}$$

for all $\mathbf{x}$ such that $\|\mathbf{x}\|_0 \leq s$.

We now provide a lower bound on the minimax risk in the case of coefficients selected according to (23) and (38).

Theorem 2. *Consider a KS dictionary learning problem with N i.i.d. observations generated according to model* (13)*. Suppose the true dictionary satisfies* (18) *for some r and fixed reference dictionary satisfying* (16)*. If the reference coordinate dictionaries $\{\mathbf{D}_{0,k}, k \in [K]\}$ satisfy* RIP$(s, \frac{1}{2})$ *and the random coefficient vector $\mathbf{x}$ is selected according to* (23) *and* (38)*, we have the following lower bound on ε^*:*

$$\varepsilon^* \geq \frac{t}{4}\min\left\{\frac{p}{s}, \frac{r^2}{2K}, \frac{\sigma^4 p}{36(3^{4K})Ns^2\sigma_a^4}\left(c_1\Big(\sum_{k\in[K]}(m_k-1)p_k\Big) - \frac{1}{2}\log_2 2K - 2\right)\right\}, \tag{40}$$

for any $0<t<1$ and any $0<c_1<\dfrac{1-t}{8\log 2}$.

Note that in Theorem 2, $\mathbf{D}$ (or its coordinate dictionaries) need not satisfy the RIP condition. Rather, the RIP is only needed for the coordinate reference dictionaries, $\{\mathbf{D}_{0,k}, k \in [K]\}$, which is a significantly weaker (and possibly trivial to satisfy) condition. We state a variation of Lemma 2 necessary for the proof of Theorem 2 — the proof is provided in the appendix.

Lemma 4. *Consider the generative model in* (13)*. Fix $r>0$ and reference dictionary $\mathbf{D}_0$ satisfying* (16)*. Then, there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{\lfloor c_1(\sum_{k\in[K]}(m_k-1)p_k) - \frac{1}{2}\log_2(2K)\rfloor}$ such that for any $0<t<1$, any $0<c_1<\frac{t^2}{8\log 2}$, any $\varepsilon'>0$ satisfying*

$$0 < \varepsilon' \leq r^2 \min\left\{\frac{1}{s}, \frac{r^2}{2Kp}\right\}, \tag{41}$$

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

$$\frac{2p}{r^2}(1-t)\varepsilon' \leq \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \leq \frac{4Kp}{r^2}\varepsilon'. \tag{42}$$

Furthermore, assuming the reference coordinate dictionaries $\{\mathbf{D}_{0,k}, k \in [K]\}$ satisfy RIP$(s, \frac{1}{2})$*, the coefficient matrix $\mathbf{X}$ is selected according to* (23) *and* (38)*, and considering side information $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$, we have:*

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq 36(3^{4K})\left(\frac{\sigma_a}{\sigma}\right)^4 \frac{Ns^2}{r^2}\varepsilon'. \tag{43}$$

Proof of Theorem 2: According to Lemma 4, for any ε' satisfying (41), there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{\lfloor c_1(\sum_{k\in[K]}(m_k-1)p_k) - \frac{K}{2}\log_2(2K)\rfloor}$ that satisfies (43) for any $0<t'<1$ and any $c_1 < \frac{t'}{8\log 2}$. Denoting $t = 1-t'$ and provided there exists an estimator with worst case MSE satisfying $\varepsilon^* \leq \frac{tp}{4}\min\{\frac{1}{s}, \frac{r^2}{2Kp}\}$, if we set $\frac{2tp}{r^2}\varepsilon' = 8\varepsilon^*$, (33) is satisfied for $\mathcal{D}_L$ and (34) holds. Consequently,

$$\frac{1}{2}\log_2(L) - 1 \leq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq \frac{36(3^{4K})}{c_2}\left(\frac{\sigma_a}{\sigma}\right)^4 \frac{Ns^2}{r^2}\varepsilon^*, \tag{44}$$

where $c_2 = \dfrac{p(1-t)}{4r^2}$. We can write (44) as

$$\varepsilon^* \geq \left(\frac{\sigma}{\sigma_a}\right)^4 \frac{tp\left(c_1\left(\sum_{k\in[K]}(m_k-1)p_k\right) - \frac{K}{2}\log_2 2K - 2\right)}{144(3^{4K})Ns^2}. \tag{45}$$

■

Focusing on the case where the coefficients follow the separable sparsity model, the next theorem provides a lower bound on the minimax risk for coefficients selected according to (24) and (38).

Theorem 3. *Consider a KS dictionary learning problem with N i.i.d. observations generated according to model* (13)*. Suppose the true dictionary satisfies* (18) *for some r and fixed reference dictionary satisfying* (16)*. If the reference coordinate dictionaries $\{\mathbf{D}_{0,k}, k \in [K]\}$ satisfy* RIP$(s, \frac{1}{2})$ *and the random coefficient vector $\mathbf{x}$ is selected according to* (24) *and* (38)*, we have the following lower bound on ε^*:*

$$\varepsilon^* \geq \frac{t}{4}\min\left\{p, \frac{r^2}{2K}, \frac{\sigma^4 p}{36(3^{4K})Ns^2\sigma_a^4}\left(c_1\Big(\sum_{k\in[K]}(m_k-1)p_k\Big) - \frac{1}{2}\log_2 2K - 2\right)\right\}, \tag{46}$$

for any $0<t<1$ and any $0<c_1<\dfrac{1-t}{8\log 2}$.

We state a variation of Lemma 4 necessary for the proof of Theorem 3. The proof of the lemma is provided in the appendix.

Lemma 5. *Consider the generative model in* (13)*. Fix $r>0$ and reference dictionary $\mathbf{D}_0$ satisfying* (16)*. Then, there exists a set of dictionaries $\mathcal{D}_L \subseteq \mathcal{D}$ of cardinality $L = 2^{\lfloor c_1(\sum_{k\in[K]}(m_k-1)p_k) - \frac{K}{2}\log_2(2K)\rfloor}$ such that for any $0<t<1$, any $0<c_1<\frac{t^2}{8\log 2}$, any $\varepsilon'>0$ satisfying*

$$0 < \varepsilon' \leq r^2 \min\left\{1, \frac{r^2}{2Kp}\right\}, \tag{47}$$

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

$$\frac{2p}{r^2}(1-t)\varepsilon' \leq \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \leq \frac{4Kp}{r^2}\varepsilon'. \tag{48}$$

Furthermore, assuming the coefficient matrix $\mathbf{X}$ is selected according to (24) *and* (38)*, the reference coordinate dictionaries $\{\mathbf{D}_{0,k}, k \in [K]\}$ satisfy* RIP$(s_k, \frac{1}{2})$*, and considering side*

information $\mathbf{T}(\mathbf{X}) = \mathrm{supp}(\mathbf{X})$*, we have:*

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq 36(3^{4K}) \left(\frac{\sigma_a}{\sigma}\right)^4 \frac{Ns^2}{r^2}\varepsilon'. \tag{49}$$

Proof of Theorem 3: The proof of Theorem 3 follows similar steps as the proof of Theorem 2. The dissimilarity arises in the condition in (47) for Lemma 5, which is different from the condition in (41) for Lemma 4. This changes the range for the minimax risk ε^* in which the lower bound in (45) holds. ∎

In the next section, we provide a simple KS dictionary learning algorithm for 2nd-order tensors and study the corresponding dictionary learning MSE.

V. Partial Converse

In the previous sections, we provided lower bounds on the minimax risk for various coefficient vector distributions and corresponding side information. We now study a special case of the problem and introduce an algorithm that achieves the lower bound in Corollary 1 (order-wise) for 2nd-order tensors. This demonstrates that our obtained lower bounds are tight in some cases.

Theorem 4. *Consider a dictionary learning problem with N i.i.d observations according to model* (13) *for $K = 2$ and let the true dictionary satisfy* (18) *for $\mathbf{D}_0 = \mathbf{I}_p$ and some $r > 0$. Further, assume the random coefficient vector $\mathbf{x}$ is selected according to* (23)*, $\mathbf{x} \in \{-1, 0, 1\}^p$, where the probabilities of the nonzero entries of $\mathbf{x}$ are arbitrary. Next, assume noise standard deviation σ and express the KS dictionary as*

$$\mathbf{D} = (\mathbf{I}_{p_1} + \mathbf{\Delta}_1) \otimes (\mathbf{I}_{p_2} + \mathbf{\Delta}_2), \tag{50}$$

where $p = p_1 p_2$, $\|\mathbf{\Delta}_1\|_F \leq r_1$ and $\|\mathbf{\Delta}_2\|_F \leq r_2$. Then, if the following inequalities are satisfied:

$$\begin{aligned} r_1\sqrt{p_2} + r_2\sqrt{p_1} + r_1 r_2 &\leq r, \\ (r_1 + r_2 + r_1 r_2)\sqrt{s} &\leq 0.1 \\ \max\left\{\frac{r_1^2}{p_2}, \frac{r_2^2}{p_1}\right\} &\leq \frac{1}{3N}, \\ \sigma &\leq 0.4, \end{aligned} \tag{51}$$

there exists a dictionary learning scheme whose MSE satisfies

$$\begin{aligned} \mathbb{E}_{\mathbf{Y}}\left\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_F^2\right\} \leq \frac{8p}{N}\left(\frac{p_1 m_1 + p_2 m_2}{m\,\mathrm{SNR}} + 3(p_1 + p_2)\right) \\ + 8p\exp\left(-\frac{0.08pN}{\sigma^2}\right), \end{aligned} \tag{52}$$

for any $\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r)$ that satisfies (50) .

To prove Theorem 4, we first introduce an algorithm to learn a KS dictionary for 2nd-order tensor data. Then, we analyze the performance of the proposed algorithm and obtain an upper bound for the MSE in the proof of Theorem 4, which is provided in the appendix.[7] Finally, we provide numerical experiments to validate our obtained results.

[7]Theorem 4 also implicitly uses the assumption that $\max\{p_1, p_2\} \leq N$.

A. KS Dictionary Learning Algorithm

We analyze a remarkably simple, two-step estimator that begins with thresholding the observations and then ends with estimating the dictionary. Note that unlike traditional dictionary learning methods, our estimator does not perform iterative alternating minimization.

a) Coefficient Estimate: We utilize a simple thresholding technique for this purpose. For all $n \in [N]$:

$$\widehat{\mathbf{x}}_n = (\widehat{x}_{n,1}, \ldots, \widehat{x}_{n,p})^\top,\ \widehat{x}_{n,l} = \begin{cases} 1 & \text{if } y_{n,l} > 0.5, \\ -1 & \text{if } y_{n,l} < -0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{53}$$

b) Dictionary Estimate: Denoting $\mathbf{A} \triangleq \mathbf{I}_{p_1} + \mathbf{\Delta}_1$ and $\mathbf{B} \triangleq \mathbf{I}_{p_2} + \mathbf{\Delta}_2$, we can write $\mathbf{D} = \mathbf{A} \otimes \mathbf{B}$. We estimate the columns of $\mathbf{A}$ and $\mathbf{B}$ separately. To learn $\mathbf{A}$, we take advantage of the Kronecker structure of the dictionary and divide each observation $\mathbf{y}_n \in \mathbb{R}^{p_1 p_2}$ into p_2 observations $\mathbf{y}'_{(n,j)} \in \mathbb{R}^{p_1}$:

$$\mathbf{y}'_{(n,j)} = \{y_{n,p_2 i + j}\}_{i=0}^{p_1 - 1},\ j \in [p_2],\ n \in [N]. \tag{54}$$

This increases the number of observations to Np_2. We also divide the original and estimated coefficient vectors:

$$\begin{aligned} \mathbf{x}'_{(n,j)} &= \{x_{n,p_2 i + j}\}_{i=0}^{p_1 - 1}, \\ \widehat{\mathbf{x}}'_{(n,j)} &= \{\widehat{x}_{n,p_2 i + j}\}_{i=0}^{p_1 - 1},\ j \in [p_2],\ n \in [N]. \end{aligned} \tag{55}$$

Similarly, we define new noise vectors:

$$\boldsymbol{\eta}'_{(n,j)} = \{\eta_{n,p_2 i + j}\}_{i=0}^{p_1 - 1},\ j \in [p_2],\ n \in [N]. \tag{56}$$

To motivate the estimation rule for the columns of $\mathbf{A}$, let us consider the original dictionary learning formulation, $\mathbf{y}_n = \mathbf{D}\mathbf{x}_n + \boldsymbol{\eta}_n$, which we can rewrite as $\mathbf{y}_n = \mathbf{x}_{n,l}\mathbf{d}_l + \sum_{i \neq l} \mathbf{x}_{n,i}\mathbf{d}_i + \boldsymbol{\eta}_n$. Multiplying both sides of the equation by $\mathbf{x}_{n,l}$ and summing up over all training data, we get $\sum_{n=1}^N \mathbf{x}_{n,l}\mathbf{y}_n = \sum_{n=1}^N (\mathbf{x}_{n,l}^2 \mathbf{d}_l + \sum_{i \neq l} \mathbf{x}_{n,l}\mathbf{x}_{n,i}\mathbf{d}_i + \mathbf{x}_{n,l}\boldsymbol{\eta}_n)$. Using the facts $\mathbb{E}_{\mathbf{x}}\{\mathbf{x}_{n,l}^2\} = \frac{s}{p}$, $\mathbb{E}_{\mathbf{x}}\{\mathbf{x}_{n,l}\mathbf{x}_{n,i}\} = 0$ for $l \neq i$, and $\mathbb{E}_{\mathbf{x},\boldsymbol{\eta}}\{\mathbf{x}_{n,l}\boldsymbol{\eta}_n\} = 0$, we get the following approximation, $\mathbf{d}_l \approx \frac{p}{Ns}\sum_{n=1}^N x_{n,l}\mathbf{y}_n$.[8] This suggests that for estimating the columns of $\mathbf{A}$, we can utilize the following equation:

$$\widetilde{\mathbf{a}}_l = \frac{p_1}{Ns}\sum_{n=1}^N \sum_{j=1}^{p_2} x'_{(k,j),l}\mathbf{y}'_{(n,j)},\ l \in [p_1]. \tag{57}$$

To estimate the columns of $\mathbf{B}$, we follow a different procedure to divide the observations. Specifically, we divide each observation $\mathbf{y}_n \in \mathbb{R}^{p_1 p_2}$ into p_1 observations $\mathbf{y}_{(n,j'')} \in \mathbb{R}^{p_2}$:

$$\mathbf{y}''_{(n,j)} = \left\{y_{n,i + p_1(j-1)}\right\}_{i=1}^{p_2},\ j \in [p_1],\ n \in [N]. \tag{58}$$

This increases the number of observations to Np_1. The coefficient vectors are also divided similarly:

$$\begin{aligned} \mathbf{x}''_{(n,j)} &= \left\{x_{k,i + p_1(j-1)}\right\}_{i=0}^{p_1 - 1}, \\ \widehat{\mathbf{x}}''_{(n,j)} &= \left\{\widehat{x}_{n,i + p_1(j-1)}\right\}_{i=0}^{p_1 - 1},\ j \in [p_1],\ n \in [N]. \end{aligned} \tag{59}$$

[8]Notice that the i.i.d. assumption on $\mathbf{x}_{n,l}$'s is critical to making this approximation work.

Similarly, we define new noise vectors:

$$\boldsymbol{\eta}''_{(n,j)} = \left\{\eta_{n,i+p_1(j-1)}\right\}_{i=1}^{p_2},\ j \in [p_1],\ n \in [N]. \tag{60}$$

Finally, using similar heuristics as the estimation rule for columns of $\mathbf{A}$, the estimate for columns of $\mathbf{B}$ can be obtained using the following equation:

$$\widetilde{\mathbf{b}}_l = \frac{p_2}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_1} x''_{(n,j),l}\mathbf{y}''_{(n,j)},\ l \in [p_2]. \tag{61}$$

The final estimate for the recovered dictionary is

$$\begin{aligned}
&\widehat{\mathbf{D}} = \widehat{\mathbf{A}} \otimes \widehat{\mathbf{B}},\\
&\widehat{\mathbf{A}} = (\widehat{\mathbf{a}}_1, \ldots, \widehat{\mathbf{a}}_{p_1}), \quad \widehat{\mathbf{a}}_l = P_{\mathcal{B}_1}(\widetilde{\mathbf{a}}_l),\\
&\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_{p_2}), \quad \widehat{\mathbf{b}}_l = P_{\mathcal{B}_1}(\widetilde{\mathbf{b}}_l),
\end{aligned} \tag{62}$$

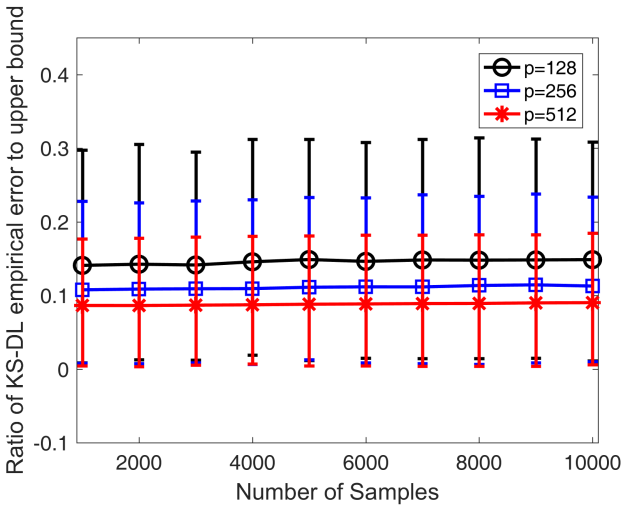
where the projection on the closed unit ball ensures that $\|\widehat{\mathbf{a}}_l\|_2 \leq 1$ and $\|\widehat{\mathbf{b}}_l\|_2 \leq 1$. Note that although projection onto the closed unit ball does not ensure the columns of $\widehat{\mathbf{D}}$ to have unit norms, our analysis only imposes this condition on the generating dictionary and the reference dictionary, and not on the recovered dictionary.

Remark 3. In addition to the heuristics following (56), the exact update rules for $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ in (57) and (61) require some additional perturbation analysis. To see this for the case of $\widetilde{\mathbf{A}}$, notice that (57) follows from writing $\mathbf{A} \otimes \mathbf{B}$ as $\mathbf{A} \otimes (\mathbf{I}_{p_2} + \boldsymbol{\Delta}_2)$, rearranging each $\mathbf{y}_n$ and $(\mathbf{A} \otimes \mathbf{I}_{p_2})\mathbf{x}_n$ into $\mathbf{y}'_{(n,j)}$'s and $\mathbf{A}\mathbf{x}'_{(n,j)}$'s, and using them to update $\widetilde{\mathbf{A}}$. In this case, we treat $(\mathbf{A} \otimes \boldsymbol{\Delta}_2)\mathbf{x}_n$ as a perturbation term in our analysis. A similar perturbation term appears in the case of the update rule for $\widetilde{\mathbf{B}}$. The analysis for dealing with these perturbation terms is provided in the appendix.

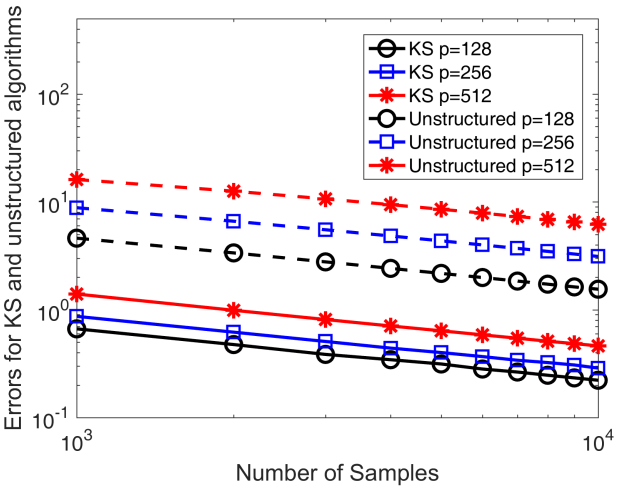
B. Empirical Comparison to Upper Bound

We are interested in empirically seeing whether our achievable scheme matches the minimax lower bound when learning KS dictionaries. To this end, we implement the preceding estimation algorithm for 2nd-order tensor data.

Figure 1(a) shows the ratio of the empirical error of the proposed KS dictionary learning algorithm in Section V-A to the obtained upper bound in Theorem 4 for 50 Monte Carlo experiments. This ratio is plotted as a function of the sample size for three choices of the number of columns p: 128, 256, and 512. The experiment shows that the ratio is approximately constant as a function of sample size, verifying the theoretical result that the estimator meets the minimax bound in terms of error scaling as a function of sample size. Figure 1(b) shows the performance of our KS dictionary learning algorithm in relation to the unstructured dictionary learning algorithm provided in [34]. It is evident that the error of our algorithm is significantly less than that for the unstructured algorithm for all three choices of p. This verifies that taking the structure of the data into consideration can indeed lead to lower dictionary identification error.



(a)



(b)

Fig. 1: Performance summary of KS dictionary learning algorithm for $p = \{128, 256, 512\}$, $s = 5$ and $r = 0.1$. (a) plots the ratio of the empirical error of our KS dictionary learning algorithm to the obtained error upper bound along with error bars for generated square KS dictionaries, and (b) shows the performance of our KS dictionary learning algorithm (solid lines) compared to the unstructured learning algorithm proposed in [34] (dashed lines).

VI. Discussion

We now discuss some of the implications of our results. Table I summarizes the lower bounds on the minimax rates from previous papers and this work. The bounds are given in terms of the number of component dictionaries K, the dictionary size parameters (m_k's and p_k's), the coefficient distribution parameters, the number of samples N, and SNR, which is defined as

$$\text{SNR} = \frac{\mathbb{E}_{\mathbf{x}}\left\{\|\mathbf{x}\|_2^2\right\}}{\mathbb{E}_{\boldsymbol{\eta}}\left\{\|\boldsymbol{\eta}\|_2^2\right\}} = \frac{\text{Tr}(\boldsymbol{\Sigma}_x)}{m\sigma^2}. \tag{63}$$

These scalings result hold for sufficiently large p and neighborhood radius r.

Comparison of minimax lower bounds for unstructured and KS dictionary learning: Compared to the results for the unstructured dictionary learning problem [34], we are able to

TABLE I: Order-wise lower bounds on the minimax risk for various coefficient distributions

Distribution \ Dictionary	Side Information $\mathbf{T}(\mathbf{X})$	Unstructured [34]	Kronecker (this paper)
1. General	$\mathbf{X}$	$\frac{\sigma^2 mp}{N\|\mathbf{\Sigma}_x\|_2}$	$\frac{\sigma^2(\sum_{k\in[K]} m_k p_k)}{NK\|\mathbf{\Sigma}_x\|_2}$
2. Sparse	$\mathbf{X}$	$\frac{p^2}{N\,\mathrm{SNR}}$	$\frac{p(\sum_{k\in[K]} m_k p_k)}{NKm\,\mathrm{SNR}}$
3. Gaussian Sparse	$\mathrm{supp}(\mathbf{X})$	$\frac{p^2}{Nm\,\mathrm{SNR}^2}$	$\frac{p(\sum_{k\in[K]} m_k p_k)}{3^{4K}Nm^2\,\mathrm{SNR}^2}$

decrease the lower bound for various coefficient distributions by reducing the scaling $\Omega(mp)$ to $\Omega(\sum_{k\in[K]} m_k p_k)$ 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 $\sum_{k\in[K]} m_k p_k$.

The results also show that the minimax risk decreases with a larger number of samples, N, and increased number of tensor order, K. By increasing K, we are shrinking the size of the class of dictionaries in which the parameter dictionary lies, thereby simplifying the problem.

Looking at the results for the general coefficient model in the first row of Table I, the lower bound for any arbitrary zero-mean random coefficient vector distribution with covariance $\mathbf{\Sigma}_x$ implies an inverse relationship between the minimax risk and SNR due to the fact that $\|\mathbf{\Sigma}_x\|_2 \leq \mathrm{Tr}(\mathbf{\Sigma}_x)$.

Comparison of general sparse and Gaussian sparse coefficient distributions: Proceeding to the sparse coefficient vector model in the second row of Table I, by replacing $\mathbf{\Sigma}_x$ with the expression in (26) in the minimax lower bound for the general coefficient distribution, we obtain the second lower bound given in (37). Recall that for s-sparse coefficient vectors,

$$\mathrm{SNR} = \frac{s\sigma_a^2}{m\sigma^2}. \tag{64}$$

Using this definition of SNR in (37), we observe a seemingly counter-intuitive increase in the MSE of order $\Omega\left(p/s\right)$ in the lower bound in comparison to the general coefficient model. However, this increase is due to the fact that we do not require coefficient vectors to have constant energy; because of this, SNR decreases for s-sparse coefficient vectors.

Next, looking at the third row of Table I, by restricting the class of sparse coefficient vector distributions to the case where non-zero elements of the coefficient vector follow a Gaussian distribution according to (38), we obtain a minimax lower bound that involves less side information than the prior two cases. However, we do make the assumption in this case that reference coordinate dictionaries satisfy $\mathsf{RIP}(s, \frac{1}{2})$. This additional assumption has two implications: (1) it introduces the factor of $1/3^{4K}$ in the minimax lower bound, and (2) it imposes the following condition on the sparsity for the "random sparsity" model: $s \leq \min_{k\in[K]}\{p_k\}$. Nonetheless, considering sparse-Gaussian coefficient vectors, we obtain a minimax lower bound that is tighter than the previous bound for some SNR values. Specifically, in order to compare bounds obtained in (37) and (40) for sparse and sparse-Gaussian coefficient vector distributions, we fix K. Then in high SNR regimes, i.e., $\mathrm{SNR} = \Omega(1/m)$, the lower bound in (37) is tighter, while (40) results in a tighter lower bound in low SNR regimes, i.e., $\mathrm{SNR} = \mathcal{O}(1/m)$, which correspond to low sparsity settings.

Comparison of random and separable sparse coefficient models: We now focus on our results for the two sparsity pattern models, namely, random sparsity and separable sparsity, for the case of sparse-Gaussian coefficient vector distribution. These results, which are reported in (40) and (46), are almost identical to each other, except for the first term in the minimization. In order to understand the settings in which the separable sparsity model in (24)—which is clearly more restrictive than the random sparsity model in (23)—turns out to be more advantageous, we select the neighborhood radius r to be of order $\mathcal{O}(\sqrt{p})$; since we are dealing with dictionaries that lie on the surface of a sphere with radius $\sqrt{p}$, this effectively ensures $\mathcal{X}(\mathbf{D}_0, r) \approx \mathcal{D}$. In this case, it can be seen from (40) and (46) that if $s = \Omega(K)$ then the separable sparsity model gives a better minimax lower bound. On the other hand, the random sparsity model should be considered for the case of $s = \mathcal{O}(K)$ because of the less restrictive nature of this model.

Achievability of our minimax lower bounds for learning KS dictionaries: To this end, we provided a simple KS dictionary learning algorithm in Section V for the special scenario of 2-dimensional tensors and analyzed the corresponding MSE, $\mathbb{E}_{\mathbf{Y}}\left\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_F^2\right\}$. In terms of scaling, the upper bound obtained for the MSE in Theorem 4 matches the lower bound in Corollary 1 provided $p_1 + p_2 < \frac{m_1 p_1 + m_2 p_2}{m\,\mathrm{SNR}}$ holds. This result suggests that more general KS dictionary learning algorithms may be developed to achieve the lower bounds reported in this paper.

VII. Conclusion

In this paper we followed an information-theoretic approach to provide lower bounds for the worst-case mean-squared error (MSE) of Kronecker-structured dictionaries that generate Kth-order tensor data. To this end, we constructed a class of Kronecker-structured dictionaries in a local neighborhood of a fixed reference Kronecker-structured dictionary. Our analysis

required studying the mutual information between the observation matrix and the dictionaries in the constructed class. To evaluate bounds on the mutual information, we considered various coefficient distributions and interrelated side information on the coefficient vectors and obtained corresponding minimax lower bounds using these models. In particular, we established that estimating Kronecker-structured dictionaries requires a number of samples that needs to grow only linearly with the sum of the sizes of the component dictionaries ($\sum_{k\in[K]} m_k p_k$), which represents the true degrees of freedom of the problem. We also demonstrated that for a special case of $K = 2$, there exists an estimator whose MSE meets the derived lower bounds. While our analysis is local in the sense that we assume the true dictionary belongs in a local neighborhood with known radius around a fixed reference dictionary, the derived minimax risk effectively becomes independent of this radius for sufficiently large neighborhood radius.

Future directions of this work include designing general algorithms to learn Kronecker-structured dictionaries that achieve the presented lower bounds. In particular, the analysis in [42] suggests that restricting the class of dictionaries to Kronecker-structured dictionaries may indeed yield a reduction in the sample complexity required for dictionary identification by replacing a factor mp in the general dictionary learning problem with the box counting dimension of the dictionary class [32].

VIII. Acknowledgement

The authors would like to thank Dr. Dionysios Kalogerias for his helpful comments.

Appendix

Proof of Lemma 1: Fix $L > 0$ and $\alpha > 0$. For a pair of matrices $\mathbf{A}_l$ and $\mathbf{A}_{l'}$, with $l \neq l'$, consider the vectorized set of entries $\mathbf{a}_l = \mathrm{vec}(\mathbf{A}_l)$ and $\mathbf{a}_{l'} = \mathrm{vec}(\mathbf{A}_{l'})$ and define the function

$$f(\mathbf{a}_l^\top, \mathbf{a}_{l'}^\top) \triangleq |\langle \mathbf{A}_l, \mathbf{A}_{l'}\rangle| = |\langle \mathbf{a}_l, \mathbf{a}_{l'}\rangle|. \tag{65}$$

For $\widetilde{\mathbf{a}} \triangleq (\mathbf{a}_l^\top, \mathbf{a}_{l'}^\top) \in \mathbb{R}^{2mp}$, write $\widetilde{\mathbf{a}} \sim \widetilde{\mathbf{a}}'$ if $\widetilde{\mathbf{a}}'$ is equal to $\widetilde{\mathbf{a}}$ in all entries but one. Then f satisfies the following bounded difference condition:

$$\sup_{\widetilde{\mathbf{a}}\sim\widetilde{\mathbf{a}}'} |f(\widetilde{\mathbf{a}}) - f(\widetilde{\mathbf{a}}')| = (\alpha - (-\alpha))\alpha = 2\alpha^2. \tag{66}$$

Hence, according to McDiarmid's inequality [43], for all $\beta > 0$, we have

$$\begin{aligned}\mathbb{P}\left(|\langle \mathbf{A}_l, \mathbf{A}_{l'}\rangle| \geq \beta\right) &\leq 2\exp\left(\frac{-2\beta^2}{\sum_{i=1}^{2mp}(2\alpha^2)^2}\right)\\ &= 2\exp\left(-\frac{\beta^2}{4\alpha^4 mp}\right).\end{aligned} \tag{67}$$

Taking a union bound over all pairs $l, l' \in [L], l \neq l'$, we have

$$\begin{aligned}&\mathbb{P}\left(\exists(l,l') \in [L]\times[L], l\neq l' : |\langle \mathbf{A}_l, \mathbf{A}_{l'}\rangle| \geq \beta\right)\\ &\qquad\leq 2L^2\exp\left(-\frac{\beta^2}{4\alpha^4 mp}\right).\end{aligned} \tag{68}$$

■

Proof of Lemma 2: Fix $r > 0$ and $t \in (0, 1)$. Let $\mathbf{D}_0$ be a reference dictionary satisfying (16), and let $\{\mathbf{U}_{(k,j)}\}_{j=1}^{p_k} \in \mathbb{R}^{m_k\times m_k}$, $k \in [K]$, be arbitrary unitary matrices satisfying

$$\mathbf{d}_{(k,0),j} = \mathbf{U}_{(k,j)}\mathbf{e}_1, \tag{69}$$

where $\mathbf{d}_{(k,0),j}$ denotes the j-th column of $\mathbf{D}_{(k,0)}$.

To construct the dictionary class $D_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$, we follow several steps. We consider sets of

$$L_k = 2^{\lfloor c_1(m_k-1)p_k - \frac{1}{2}\log_2 2K\rfloor} \tag{70}$$

generating matrices $\mathbf{G}_{(k,l_k)}$:

$$\mathbf{G}_{(k,l_k)} \in \left\{-\frac{1}{r^{1/K}\sqrt{(m_k-1)}}, \frac{1}{r^{1/K}\sqrt{(m_k-1)}}\right\}^{(m_k-1)\times p_k} \tag{71}$$

for $k \in [K]$ and $l_k \in [L_k]$. According to Lemma 1, for all $k \in [K]$ and any $\beta > 0$, the following relation is satisfied:

$$\begin{aligned}&\mathbb{P}\left(\exists(l_k, l_k') \in [L_k]\times[L_k], l \neq l' : \left|\left\langle \mathbf{G}_{(k,l_k)}, \mathbf{G}_{(k,l_k')}\right\rangle\right| \geq \beta\right)\\ &\qquad\leq 2L_k^2\exp\left(-\frac{r^{4/K}(m_k-1)\beta^2}{4p_k}\right).\end{aligned} \tag{72}$$

To guarantee a simultaneous existence of K sets of generating matrices satisfying

$$\left|\left\langle \mathbf{G}_{(k,l_k)}, \mathbf{G}_{(k,l_k')}\right\rangle\right| \leq \beta, \quad k \in [K], \tag{73}$$

we take a union bound of (72) over all $k \in [K]$ and choose parameters such that the following upper bound is less than 1:

$$\begin{aligned}&2KL_k^2\exp\left(-\frac{r^{4/K}(m_k-1)\beta^2}{4p_k}\right)\\ &\qquad= \exp\left(-\frac{r^{4/K}(m_k-1)\beta^2}{4p_k} + 2\ln\sqrt{2K}L_k\right),\end{aligned} \tag{74}$$

which is satisfied as long as the following inequality holds:

$$\log_2 L_k < \frac{r^{4/K}(m_k-1)\beta^2}{8p_k\log 2} - \frac{1}{2} - \frac{1}{2}\log_2 K. \tag{75}$$

Now, setting $\beta = \frac{p_k t}{r^{2/K}}$, the condition in (75) holds and there exists a collection of generating matrices that satisfy:

$$\left|\left\langle \mathbf{G}_{(k,l_k)}, \mathbf{G}_{(k,l_k')}\right\rangle\right| \leq \frac{p_k t}{r^{2/K}}, \quad k \in [K], \tag{76}$$

for any distinct $l_k, l_k' \in [L_k]$, any $t \in (0, 1)$, and any $c_1 > 0$ such that

$$c_1 < \frac{t^2}{8\log 2}. \tag{77}$$

We next construct matrices that will be later used for the construction of unit-norm column dictionaries. We construct $\mathbf{D}_{(k,1,l_k)} \in \mathbb{R}^{m_k\times p_k}$ column-wise using $\mathbf{G}_{(k,l_k)}$ and unitary matrices $\{\mathbf{U}_{(k,j)}\}_{j=1}^{p_k}$. Let the j-th column of $\mathbf{D}_{(k,1,l_k)}$ be given by

$$\mathbf{d}_{(k,1,l_k),j} = \mathbf{U}_{(k,j)}\begin{pmatrix}0\\ \mathbf{g}_{(k,l_k),j}\end{pmatrix}, \quad k \in [K], \tag{78}$$

for any $l_k \in [L_k]$. Moreover, defining

$$\mathcal{D}_1 \triangleq \left\{ \bigotimes_{k\in[K]} \mathbf{D}_{(k,1,l_k)} : l_k \in [L_k] \right\}, \tag{79}$$

and denoting

$$\mathcal{L} \triangleq \{(l_1, \ldots, l_K) : l_k \in [L_k]\}, \tag{80}$$

any element of $\mathcal{D}_1$ can be expressed as

$$\mathbf{D}_{(1,l)} = \bigotimes_{k\in[K]} \mathbf{D}_{(k,1,l_k)}, \forall\ l \in [L], \tag{81}$$

where $|\mathcal{L}| = L \triangleq \prod_{k\in[K]} L_k$ and we associate an $l \in [L]$ with a tuple in $\mathcal{L}$ via lexicographic indexing. Notice also that

$$\left\|\mathbf{d}_{(1,l),j}\right\|_2^2 \overset{(a)}{=} \prod_{k\in[K]} \left\|\mathbf{d}_{(k,1,l_k),j}\right\|_2^2 = \prod_{k\in[K]} \frac{1}{r^{2/K}} = \frac{1}{r^2}, \text{ and}$$
$$\left\|\mathbf{D}_{(1,l)}\right\|_F^2 = \frac{p}{r^2}, \tag{82}$$

where (a) follows from properties of the Kronecker product. From (78), it is evident that for all $k \in [K]$, $\mathbf{d}_{(k,0),j}$ is orthogonal to $\mathbf{d}_{(k,1,l_k),j}$ and consequently, we have

$$\left\langle \mathbf{D}_{(k,0)}, \mathbf{D}_{(k,1,l_k)} \right\rangle = 0,\ k \in [K] \tag{83}$$

Also,

$$\begin{aligned}
\left\langle \mathbf{D}_{(k,1,l_k)}, \mathbf{D}_{(k,1,l'_k)} \right\rangle &= \sum_{j=1}^{p_k} \left\langle \mathbf{d}_{(k,1,l_k),j}, \mathbf{d}_{(k,1,l'_k),j} \right\rangle \\
&= \sum_{j=1}^{p_k} \left\langle \mathbf{U}_{(k,j)} \begin{pmatrix} 0 \\ \mathbf{g}_{(k,l_k),j} \end{pmatrix}, \mathbf{U}_{(k,j)} \begin{pmatrix} 0 \\ \mathbf{g}_{(k,l'_k),j} \end{pmatrix} \right\rangle \\
&\overset{(b)}{=} \sum_{j=1}^{p_k} \left\langle \mathbf{g}_{(k,l_k),j}, \mathbf{g}_{(k,l'_k),j} \right\rangle \\
&= \left\langle \mathbf{G}_{(k,l_k)}, \mathbf{G}_{(k,l'_k)} \right\rangle,
\end{aligned} \tag{84}$$

where (b) follows from the fact that $\{\mathbf{U}_{(k,j)}\}$ are unitary.

Based on the construction, for all $k \in [K]$, $l_k, l'_k \in [L_k]$, $l_k \neq l'_k$, we have

$$\begin{aligned}
&\left\|\mathbf{D}_{(1,l)} - \mathbf{D}_{(1,l')}\right\|_F^2 \\
&\quad = \left\|\mathbf{D}_{(1,l)}\right\|_F^2 + \left\|\mathbf{D}_{(1,l')}\right\|_F^2 - 2\left\langle \mathbf{D}_{(1,l)}, \mathbf{D}_{(1,l')} \right\rangle \\
&\quad = \frac{p}{r^2} + \frac{p}{r^2} - 2\prod_{k\in[K]} \left\langle \mathbf{D}_{(k,1,l_k)}, \mathbf{D}_{(k,1,l'_k)} \right\rangle \\
&\quad \geq 2\left(\frac{p}{r^2} - \prod_{k\in[K]} \left| \left\langle \mathbf{D}_{(k,1,l_k)}, \mathbf{D}_{(k,1,l'_k)} \right\rangle \right| \right) \\
&\quad \overset{(c)}{=} 2\left(\frac{p}{r^2} - \prod_{k\in[K]} \left| \left\langle \mathbf{G}_{(k,l_k)}, \mathbf{G}_{(k,l'_k)} \right\rangle \right| \right) \\
&\quad \overset{(d)}{\geq} 2\left(\frac{p}{r^2} - \prod_{k\in[K]} \frac{p_k}{r^{2/K}} t \right) \\
&\quad = \frac{2p}{r^2}\left(1 - t^K\right),
\end{aligned} \tag{85}$$

where (c) and (d) follow from (84) and (76), respectively.

We are now ready to define $\mathcal{D}_L$. The final dictionary class is defined as

$$\mathcal{D}_L \triangleq \left\{ \bigotimes_{k\in[K]} \mathbf{D}_{(k,l_k)} : l_k \in [L_k] \right\} \tag{86}$$

and any $\mathbf{D}_l \in \mathcal{D}_L$ can be written as

$$\mathbf{D}_l = \bigotimes_{k\in[K]} \mathbf{D}_{(k,l_k)}, \tag{87}$$

where $\mathbf{D}_{(k,l_k)}$ is defined as

$$\mathbf{D}_{(k,l_k)} \triangleq \eta \mathbf{D}_{(k,0)} + \nu \mathbf{D}_{(k,1,l_k)}, \quad k \in [K], \tag{88}$$

and

$$\eta \triangleq \sqrt{1 - \frac{\varepsilon'}{r^2}}, \quad \nu \triangleq \sqrt{\frac{r^{2/K}\varepsilon'}{r^2}} \tag{89}$$

for any

$$0 < \varepsilon' < \min\left\{ r^2, \frac{r^4}{2Kp} \right\}, \tag{90}$$

which ensures that $1 - \frac{\varepsilon'}{r^2} > 0$ and $\mathbf{D}_l \in \mathcal{X}(\mathbf{D}_0, r)$. Note that the following relation holds between η and ν:

$$\eta^2 + \frac{\nu^2}{r^{2/K}} = 1. \tag{91}$$

We can expand (87) to facilitate the forthcoming analysis:

$$\mathbf{D}_l = \sum_{\mathbf{i}\in\{0,1\}^K} \eta^{K-\|\mathbf{i}\|_1} \nu^{\|\mathbf{i}\|_1} \left(\bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l_k)} \right), \tag{92}$$

where $\mathbf{i} \triangleq (i_1, i_2, \ldots, i_K)$ and $\mathbf{D}_{(k,0,l_k)} \triangleq \mathbf{D}_{(k,0)}$. To show $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$, we first show that any $\mathbf{D}_l \in \mathcal{D}_L$ has unit-norm columns. For any $j \in [p]$ and $j_k \in [p_k], k \in [K]$ (associating j with $(j_1, \ldots, j_K)$ via lexicographic indexing), we have

$$\begin{aligned}
\|\mathbf{d}_{l,j}\|_2^2 &= \prod_{k\in[K]} \left\|\mathbf{d}_{(k,l_k),j_k}\right\|_2^2 \\
&= \prod_{k\in[K]} \left(\eta^2 \|\mathbf{d}_{(k,0),j_k}\|_2^2 + \nu^2 \left\|\mathbf{d}_{(k,1,l_k),j_k}\right\|_2^2 \right) \\
&= \prod_{k\in[K]} \left(\eta^2 + \nu^2 \left(\frac{1}{r^{2/K}}\right) \right) \\
&\overset{(e)}{=} 1,
\end{aligned} \tag{93}$$

where (e) follows from (91). Then, we show that $\|\mathbf{D}_l - \mathbf{D}_0\|_F \leq r$:

$$\begin{aligned}
&\|\mathbf{D}_l - \mathbf{D}_0\|_F^2 \\
&= \left\| \mathbf{D}_0 - \sum_{\mathbf{i}\in\{0,1\}^K} \eta^{K-\|\mathbf{i}\|_1} \nu^{\|\mathbf{i}\|_1} \bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l_k)} \right\|_F^2 \\
&= \left\| \left(1 - \eta^K\right)\mathbf{D}_0 - \sum_{\substack{\mathbf{i}\in\{0,1\}^K \\ \|\mathbf{i}\|_1 \neq 0}} \eta^{K-\|\mathbf{i}\|_1} \nu^{\|\mathbf{i}\|_1} \bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l_k)} \right\|_F^2 \\
&= \left(1 - \eta^K\right)^2 \|\mathbf{D}_0\|_F^2
\end{aligned}$$

$$+\sum_{\substack{\mathbf{i}\in\{0,1\}^K\\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)}\nu^{2\|\mathbf{i}\|_1}\prod_{k\in[K]}\left\|\mathbf{D}_{(k,i_k,l_k)}\right\|_F^2. \quad (94)$$

We will bound the two terms in (94) separately. We know

$$(1-x^n)=(1-x)(1+x+x^2+\cdots+x^{n-1}). \quad (95)$$

Hence, we have

$$\begin{aligned}\left(1-\eta^K\right)^2\|\mathbf{D}_0\|_F^2 &= \left(1-\eta^K\right)^2 p\\ &\overset{(f)}{\leq} \left(1-\eta^K\right)p\\ &\leq \left(1-\eta^{2K}\right)p\\ &\overset{(g)}{=} \left(1-\eta^2\right)\left(1+\eta^2+\cdots+\eta^{2(K-1)}\right)p\\ &= \frac{\varepsilon'}{r^2}\left(1+\eta^2+\cdots+\eta^{2(K-1)}\right)p\\ &\overset{(h)}{\leq} \frac{Kp\varepsilon'}{r^2}, \end{aligned} \quad (96)$$

where (f) and (h) follow from the fact that $\eta<1$ and (g) follows from (95).

Similarly for the second term in (94),

$$\begin{aligned}&\prod_{k\in[K]}\left\|\mathbf{D}_{(k,i_k,l_k)}\right\|_F^2\\ &\quad=\left(\prod_{\substack{k\in[K]\\ i_k=0}}\|\mathbf{D}_{(k,0)}\|_F^2\right)\left(\prod_{\substack{k\in[K]\\ i_k=1}}\|\mathbf{D}_{(k,1,l_k)}\|_F^2\right)\\ &\quad=\left(\prod_{\substack{k\in[K]\\ i_k=0}}p_k\right)\left(\prod_{\substack{k\in[K]\\ i_k=1}}\frac{p_k}{r^{2/K}}\right)\\ &\quad=\left(\prod_{k\in[K]}p_k\right)\left(\frac{1}{r^{2/K}}\right)^{\|\mathbf{i}\|_1}. \end{aligned} \quad (97)$$

Replacing values for η and ν from (89) and using (97) and the fact that $\prod_{k\in[K]}p_k=p$, we can further reduce the second term in (94) to get

$$\begin{aligned}&\sum_{\substack{\mathbf{i}\in\{0,1\}^K\\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)}\nu^{2\|\mathbf{i}\|_1}\prod_{k\in[K]}\|\mathbf{D}_{(k,i_k,l_k)}\|_F^2\\ &\quad= p\sum_{k=0}^{K-1}\binom{K}{k}\left(1-\frac{\varepsilon'}{r^2}\right)^k\left(\frac{\varepsilon'}{r^2}\right)^{K-k}\\ &\quad= p\left(1-\left(1-\frac{\varepsilon'}{r^2}\right)^K\right)\\ &\quad\overset{(i)}{=} p\left(\frac{\varepsilon'}{r^2}\right)\left(1+\left(1-\frac{\varepsilon'}{r^2}\right)+\cdots+\left(1-\frac{\varepsilon'}{r^2}\right)^{K-1}\right)\\ &\quad\leq \frac{Kp\varepsilon'}{r^2}, \end{aligned} \quad (98)$$

where (i) follows from (95). Adding (96) and (98), we get

$$\begin{aligned}\|\mathbf{D}_l-\mathbf{D}_0\|_F^2 &\leq \varepsilon'\left(\frac{2Kp}{r^2}\right)\\ &\overset{(j)}{\leq} r^2, \end{aligned} \quad (99)$$

where (j) follows from the condition in (90). Therefore, (93) and (98) imply that $\mathcal{D}_L\subseteq\mathcal{X}(\mathbf{D}_0,r)$.

We now find lower and upper bounds for the distance between any two distinct elements $\mathbf{D}_l,\mathbf{D}_{l'}\in\mathcal{D}_L$.

1) Lower bounding $\|\mathbf{D}_l-\mathbf{D}_{l'}\|_F^2$: We define the set $\mathcal{I}_i\subseteq[K]$ where $|\mathcal{I}_i|=i, i\in[K]$. Then, given distinct $l_k,l_k',k\in\mathcal{I}_i$, we have

$$\begin{aligned}\left\|\bigotimes_{k\in\mathcal{I}_i}\mathbf{D}_{(k,1,l_k)}-\bigotimes_{k\in\mathcal{I}_i}\mathbf{D}_{(k,1,l_k')}\right\|_F^2 &\overset{(k)}{\geq} \frac{2\left(1-t^i\right)}{r^{2i/K}}\prod_{k\in\mathcal{I}_i}p_k\\ &\geq \frac{2\left(1-t\right)}{r^{2i/K}}\prod_{k\in\mathcal{I}_i}p_k, \end{aligned} \quad (100)$$

where (k) follows using arguments similar to those made for (85).

To obtain a lower bound on $\|\mathbf{D}_l-\mathbf{D}_{l'}\|_F^2$, we emphasize that for distinct $l,l'\in[L]$, it does not necessarily hold that $l_k\neq l_k'$ for all $k\in[K]$. In fact, it is sufficient for $\mathbf{D}_l\neq\mathbf{D}_{l'}$ that only one $k\in[K]$ satisfies $l_k\neq l_k'$. Now, assume only K_1 out of K coordinate dictionaries are distinct (for the case where all smaller dictionaries are distinct, $K_1=K$). Without loss of generality, we assume $l_1,\ldots,l_{K_1}$ are distinct and $l_{K_1+1},\ldots,l_K$ are identical across $\mathbf{D}_l$ and $\mathbf{D}_{l'}$. This is because of the invariance of the Frobenius norm of Kronecker products under permutation, i.e.,

$$\left\|\bigotimes_{k\in[K]}\mathbf{A}_k\right\|_F=\prod_{k\in[K]}\|\mathbf{A}_k\|_F=\left\|\bigotimes_{k\in[K]}\mathbf{A}_{\pi(k)}\right\|_F, \quad (101)$$

where $\pi(.)$ denotes a permutation of $[K]$. We then have

$$\begin{aligned}&\|\mathbf{D}_l-\mathbf{D}_{l'}\|_F^2\\ &\quad=\Big\|(\mathbf{D}_{(1,l_1)}\otimes\cdots\otimes\mathbf{D}_{(K_1,l_{K_1})}\otimes\\ &\qquad\mathbf{D}_{(K_1+1,l_{K_1+1})}\otimes\cdots\otimes\mathbf{D}_{(K,l_K)})\\ &\qquad-(\mathbf{D}_{(1,l_1')}\otimes\cdots\otimes\mathbf{D}_{(K_1,l_{K_1}')}\otimes\\ &\qquad\mathbf{D}_{(K_1+1,l_{K_1+1})}\otimes\cdots\otimes\mathbf{D}_{(K,l_K)})\Big\|_F^2\\ &\quad\overset{(l)}{=}\Big\|\Big(\bigotimes_{k\in[K_1]}\mathbf{D}_{(k,l_k)}-\bigotimes_{k\in[K_1]}\mathbf{D}_{(k,l_k')}\Big)\otimes\\ &\qquad\mathbf{D}_{(K_1+1,l_{K_1+1})}\otimes\cdots\otimes\mathbf{D}_{(K,l_K)})\Big\|_F^2\\ &\quad=\left\|\bigotimes_{k\in[K_1]}\mathbf{D}_{(k,l_k)}-\bigotimes_{k\in[K_1]}\mathbf{D}_{(k,l_k')}\right\|_F^2\prod_{k=K_1+1}^{K}\|\mathbf{D}_{(k,l_k)}\|_F^2\\ &\quad=\left(\prod_{k=K_1+1}^{K}p_k\right)\Big\|\sum_{\substack{\mathbf{i}\in\{0,1\}^{K_1}\\ \|\mathbf{i}\|_1\neq 0}}\eta^{K_1-\|\mathbf{i}\|_1}\nu^{\|\mathbf{i}\|_1}\\ &\qquad\left(\bigotimes_{k\in[K_1]}\mathbf{D}_{(k,i_k,l_k)}-\bigotimes_{k\in[K_1]}\mathbf{D}_{(k,i_k,l_k')}\right)\Big\|_F^2\\ &\quad\overset{(m)}{=}\Bigg(\sum_{\substack{\mathbf{i}\in\{0,1\}^{K_1}\\ \|\mathbf{i}\|_1\neq 0}}\eta^{2(K_1-\|\mathbf{i}\|_1)}\nu^{2\|\mathbf{i}\|_1}\prod_{\substack{k\in[K_1]\\ i_k=0}}\left\|\mathbf{D}_{(k,0)}\right\|_F^2 \end{aligned}$$

$$\left\| \bigotimes_{\substack{k\in[K_1]\\ i_k=1}} \mathbf{D}_{(k,1,l_k)} - \bigotimes_{\substack{k\in[K_1]\\ i_k=1}} \mathbf{D}_{(k,1,l'_k)} \right\|_F^2 \Bigg)$$

$$\overset{(n)}{\geq} \left(\prod_{k=K_1+1}^{K} p_k \right) \Bigg(\sum_{\substack{\mathbf{i}\in\{0,1\}^{K_1}\\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K_1-\|\mathbf{i}\|_1)} \nu^{2\|\mathbf{i}\|_1} \left(\prod_{\substack{k\in[K_1]\\ i_k=0}} p_k \right) \left(\frac{2}{r^{2\|\mathbf{i}\|_1/K}} \prod_{\substack{k\in[K_1]\\ i_k=1}} p_k \right) (1-t) \Bigg)$$

$$\overset{(o)}{=} 2p\,(1-t) \sum_{k=0}^{K_1-1} \binom{K_1}{k} \left(1-\frac{\varepsilon'}{r^2}\right)^k \left(\frac{\varepsilon'}{r^2}\right)^{K_1-k}$$

$$\overset{(p)}{=} 2p\,(1-t) \left(1 - \left(1-\frac{\varepsilon'}{r^2}\right)^{K_1} \right)$$

$$\geq 2p\,(1-t) \left(1 - \left(1-\frac{\varepsilon'}{r^2}\right) \right)$$

$$= \frac{2p}{r^2}\,(1-t)\,\varepsilon', \tag{102}$$

where (l) follows from the distributive property of Kronecker products, (m) follows the fact that terms in the sum have orthogonal columns (from (4) and (83)), (n) follows from (100), (o) follows from substituting values for η and ν, and (p) follows from the binomial formula.

2) Upper bounding $\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2$: In order to upper bound $\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2$, notice that

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2$$

$$= \sum_{\substack{\mathbf{i}\in\{0,1\}^K\\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)} \nu^{2\|\mathbf{i}\|_1} \left\| \bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l_k)} - \bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l'_k)} \right\|_F^2$$

$$\overset{(q)}{\leq} \sum_{\substack{\mathbf{i}\in\{0,1\}^K\\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)} \nu^{2\|\mathbf{i}\|_1} \left(\left\| \bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l_k)} \right\|_F + \left\| \bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l'_k)} \right\|_F \right)^2$$

$$= 4 \sum_{\substack{\mathbf{i}\in\{0,1\}^K\\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)} \nu^{2\|\mathbf{i}\|_1} \left\| \bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l_k)} \right\|_F^2$$

$$= 4 \sum_{\substack{\mathbf{i}\in\{0,1\}^K\\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)} \nu^{2\|\mathbf{i}\|_1} \prod_{\substack{k\in[K]\\ i_k=0}} \|\mathbf{D}_{(k,0)}\|_F^2 \prod_{\substack{k\in[K]\\ i_k=1}} \|\mathbf{D}_{(k,1,l_k)}\|_F^2$$

$$= 4 \sum_{\substack{\mathbf{i}\in\{0,1\}^K\\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)} \nu^{2\|\mathbf{i}\|_1} \left(\prod_{\substack{k\in[K]\\ i_k=0}} p_k \right) \left(\prod_{\substack{k\in[K]\\ i_k=1}} \frac{p_k}{r^{2/K}} \right)$$

$$\overset{(r)}{=} 4p \sum_{k=0}^{K-1} \binom{K}{k} \left(1-\frac{\varepsilon'}{r^2}\right)^k \left(\frac{\varepsilon'}{r^2}\right)^{K-k}$$

$$\overset{(s)}{\leq} \frac{4Kp}{r^2}\varepsilon', \tag{103}$$

where (q) follows from the triangle inequality, (r) follows from substituting values for η and ν, and (s) follows from similar arguments as in (98).

3) Upper bounding $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$: We next obtain an upper bound for $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$ for the dictionary set $\mathcal{D}_L$ according to the general coefficient model and side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$.

Assuming side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, conditioned on the coefficients $\mathbf{x}_n$, the observations $\mathbf{y}_n$ follow a multivariate Gaussian distribution with covariance matrix $\sigma^2\mathbf{I}$ and mean vector $\mathbf{D}\mathbf{x}_n$. From the convexity of the KL divergence [44], following similar arguments as in [34], [40], we have

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) = I(\mathbf{Y}; l|\mathbf{X})$$

$$= \frac{1}{L} \sum_{l\in[L]} \mathbb{E}_{\mathbf{X}} \left\{ D_{KL} \left(f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X}) \middle\| \frac{1}{L} \sum_{l'\in[L]} f_{\mathbf{D}_{l'}}(\mathbf{Y}|\mathbf{X}) \right) \right\}$$

$$\leq \frac{1}{L^2} \sum_{l,l'\in[L]} \mathbb{E}_{\mathbf{X}} \left\{ D_{KL} \left(f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X}) \middle\| f_{\mathbf{D}_{l'}}(\mathbf{Y}|\mathbf{X}) \right) \right\}, \tag{104}$$

where $f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X})$ is the probability distribution of the observations $\mathbf{Y}$, given the coefficient matrix $\mathbf{X}$ and the dictionary $\mathbf{D}_l$. From Durrieu et al. [45], we have

$$D_{KL} \left(f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X}) \middle\| f_{\mathbf{D}_{l'}}(\mathbf{Y}|\mathbf{X}) \right)$$

$$= \sum_{n\in[N]} \frac{1}{2\sigma^2} \|(\mathbf{D}_l - \mathbf{D}_{l'})\mathbf{x}_n\|_2^2$$

$$= \sum_{n\in[N]} \frac{1}{2\sigma^2} \operatorname{Tr}\left\{ (\mathbf{D}_l - \mathbf{D}_{l'})^\top (\mathbf{D}_l - \mathbf{D}_{l'}) \mathbf{x}_n \mathbf{x}_n^\top \right\}. \tag{105}$$

Substituting (105) in (104) results in

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$$

$$\leq \mathbb{E}_{\mathbf{X}} \left\{ \sum_{n\in[N]} \frac{1}{2\sigma^2} \operatorname{Tr}\left\{ (\mathbf{D}_l - \mathbf{D}_{l'})^\top (\mathbf{D}_l - \mathbf{D}_{l'}) \mathbf{x}_n \mathbf{x}_n^\top \right\} \right\}$$

$$= \sum_{n\in[N]} \frac{1}{2\sigma^2} \operatorname{Tr}\left\{ (\mathbf{D}_l - \mathbf{D}_{l'})^\top (\mathbf{D}_l - \mathbf{D}_{l'}) \mathbf{\Sigma}_x \right\}$$

$$\overset{(t)}{\leq} \sum_{n\in[N]} \frac{1}{2\sigma^2} \|\mathbf{\Sigma}_x\|_2 \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2$$

$$\overset{(u)}{\leq} \frac{N}{2\sigma^2} \|\mathbf{\Sigma}_x\|_2 \left(\frac{4Kp\varepsilon'}{r^2} \right)$$

$$= \frac{2NKp\|\mathbf{\Sigma}_x\|_2}{r^2\sigma^2}\varepsilon', \tag{106}$$

where (u) follows from (103). To show (t), we use the fact that for any $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\mathbf{\Sigma}_x$ with ordered singular values $\sigma_i(\mathbf{A})$ and $\sigma_i(\mathbf{\Sigma}_x), i \in [p]$, we have

$$\operatorname{Tr}\{\mathbf{A}\mathbf{\Sigma}_x\} \leq |\operatorname{Tr}\{\mathbf{A}\mathbf{\Sigma}_x\}|$$

$$\overset{(v)}{\leq} \sum_{i=1}^{p} \sigma_i(\mathbf{A})\sigma_i(\mathbf{\Sigma}_x)$$

$$\overset{(w)}{\leq} \sigma_1(\mathbf{\Sigma}_x)\sum_{i=1}^{p}\sigma_i(\mathbf{A}) = \|\mathbf{\Sigma}_x\|_2 \operatorname{Tr}\{\mathbf{A}\}, \tag{107}$$

where (v) follows from Von Neumann's trace inequality [46] and (w) follows from the positivity of the singular values of $\mathbf{\Sigma}_x$. The inequality in (t) follows from replacing $\mathbf{A}$ with $(\mathbf{D}_l-\mathbf{D}_{l'})^\top(\mathbf{D}_l-\mathbf{D}_{l'})$ and using the fact that $\operatorname{Tr}\{(\mathbf{D}_l-\mathbf{D}_{l'})^\top(\mathbf{D}_l-\mathbf{D}_{l'})\} = \|\mathbf{D}_l-\mathbf{D}_{l'}\|_F^2$. ∎

Proof of Lemma 4: The dictionary class $\mathcal{D}_L$ constructed in Lemma 2 is again considered here. Note that (41) implies $\varepsilon' < r^2$, since $s \geq 1$. The first part of Lemma 4, up to (42), thus trivially follows from Lemma 2. In order to prove the second part, notice that in this case the coefficient vector is assumed to be sparse according to (23). Denoting $\mathbf{x}_{\mathcal{S}_n}$ as the elements of $\mathbf{x}_n$ with indices $\mathcal{S}_n \triangleq \operatorname{supp}(\mathbf{x}_n)$, we have observations $\mathbf{y}_n$ as

$$\mathbf{y}_n = \mathbf{D}_{l,\mathcal{S}_n}\mathbf{x}_{\mathcal{S}_n} + \boldsymbol{\eta}_n. \tag{108}$$

Hence conditioned on $\mathcal{S}_n = \operatorname{supp}(\mathbf{x}_n)$, observations $\mathbf{y}_n$'s are zero-mean independent multivariate Gaussian random vectors with covariances

$$\mathbf{\Sigma}_{(n,l)} = \sigma_a^2\mathbf{D}_{l,\mathcal{S}_n}\mathbf{D}_{l,\mathcal{S}_n}^\top + \sigma^2\mathbf{I}_s. \tag{109}$$

The conditional MI $I(\mathbf{Y};l|\mathbf{T}(\mathbf{X}) = \operatorname{supp}(\mathbf{X}))$ has the following upper bound [34], [47]:

$$\begin{aligned} I(\mathbf{Y};l|\mathbf{T}(\mathbf{X})) &\leq \mathbb{E}_{\mathbf{T}(\mathbf{X})}\Bigg\{ \sum_{\substack{n\in[N]\\ l,l'\in[L]}} \frac{1}{L^2} \\ &\qquad \operatorname{Tr}\left\{\left[\mathbf{\Sigma}_{(n,l)}^{-1}-\mathbf{\Sigma}_{(n,l')}^{-1}\right]\left[\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}\right]\right\}\Bigg\} \\ &\leq \operatorname{rank}\left\{\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}\right\}\mathbb{E}_{\mathbf{T}(\mathbf{X})}\Bigg\{ \sum_{n\in[N]}\frac{1}{L^2} \\ &\qquad \sum_{l,l'\in[L]}\left\|\mathbf{\Sigma}_{(n,l)}^{-1}-\mathbf{\Sigma}_{(n,l')}^{-1}\right\|_2\left\|\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}\right\|_2\Bigg\}. \end{aligned} \tag{110}$$

Since $\operatorname{rank}(\mathbf{\Sigma}_{(n,l)}) \leq s$, $\operatorname{rank}\{\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}\} \leq 2s$ [34].

Next, note that since non-zero elements of the coefficient vector are selected according to (23) and (38), we can write the subdictionary $\mathbf{D}_{l,\mathcal{S}_n}$ in terms of the Khatri-Rao product of matrices:

$$\mathbf{D}_{l,\mathcal{S}_n} = \underset{k\in[K]}{\circledast}\mathbf{D}_{(k,l_k),\mathcal{S}_{n_k}}, \tag{111}$$

where $\mathcal{S}_{n_k} = \{j_{n_k}\}_{n_k=1}^s, j_{n_k} \in [p_k]$, for any $k \in [K]$, denotes the support of $\mathbf{x}_n$ according to the coordinate dictionary $\mathbf{D}_{(k,l_k)}$ and $\mathcal{S}_n$ corresponds to the indexing of the elements of $(\mathcal{S}_1 \times \ldots \mathcal{S}_K)$. Note that $\mathbf{D}_{l,\mathcal{S}_n} \in \mathbb{R}^{(\prod_{k\in[K]} m_k)\times s}$ and in this case, the $\mathcal{S}_{n_k}$'s can be multisets.[9] We can now write

$$\mathbf{\Sigma}_{(n,l)} = \sigma_a^2\left(\underset{k_1\in[K]}{\circledast}\mathbf{D}_{(k_1,l_{k_1}),\mathcal{S}_{n_{k_1}}}\right)\left(\underset{k_2\in[K]}{\circledast}\mathbf{D}_{(k_2,l_{k_2}),\mathcal{S}_{n_{k_2}}}\right)^\top + \sigma^2\mathbf{I}_s. \tag{112}$$

We next write

$$\begin{aligned} &\frac{1}{\sigma_a^2}(\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}) \\ &= \left(\underset{k_1\in[K]}{\circledast}\mathbf{D}_{(k_1,l_{k_1}),\mathcal{S}_{n_{k_1}}}\right)\left(\underset{k_2\in[K]}{\circledast}\mathbf{D}_{(k_2,l_{k_2}),\mathcal{S}_{n_{k_2}}}\right)^\top \\ &\quad -\left(\underset{k_1\in[K]}{\circledast}\mathbf{D}_{(k_1,l'_{k_1}),\mathcal{S}_{n_{k_1}}}\right)\left(\underset{k_2\in[K]}{\circledast}\mathbf{D}_{(k_2,l'_{k_2}),\mathcal{S}_{n_{k_2}}}\right)^\top \\ &= \left(\sum_{\mathbf{i}\in\{0,1\}^K}\eta^{K-\|\mathbf{i}\|_1}\nu^{\|\mathbf{i}\|_1}\underset{k_1\in[K]}{\circledast}\mathbf{D}_{(k_1,i_{k_1},l_{k_1}),\mathcal{S}_{n_{k_1}}}\right) \\ &\qquad \left(\sum_{\mathbf{i}'\in\{0,1\}^K}\eta^{K-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}'\|_1}\underset{k_2\in[K]}{\circledast}\mathbf{D}_{(k_2,i'_{k_2},l_{k_2}),\mathcal{S}_{n_{k_2}}}\right)^\top \\ &\quad -\left(\sum_{\mathbf{i}\in\{0,1\}^K}\eta^{K-\|\mathbf{i}\|_1}\nu^{\|\mathbf{i}\|_1}\underset{k_1\in[K]}{\circledast}\mathbf{D}_{(k_1,i_{k_1},l'_{k_1}),\mathcal{S}_{n_{k_1}}}\right) \\ &\qquad \left(\sum_{\mathbf{i}'\in\{0,1\}^K}\eta^{K-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}'\|_1}\underset{k_2\in[K]}{\circledast}\mathbf{D}_{(k_2,i'_{k_2},l'_{k_2}),\mathcal{S}_{n_{k_2}}}\right)^\top \\ &= \sum_{\substack{\mathbf{i},\mathbf{i}'\in\{0,1\}^K\\ \|\mathbf{i}\|_1+\|\mathbf{i}'\|_1\neq 0}}\eta^{2K-\|\mathbf{i}\|_1-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}\|_1+\|\mathbf{i}'\|_1} \\ &\qquad \left(\underset{k_1\in[K]}{\circledast}\mathbf{D}_{(k_1,i_{k_1},l_{k_1}),\mathcal{S}_{n_{k_1}}}\right)\left(\underset{k_2\in[K]}{\circledast}\mathbf{D}_{(k_2,i'_{k_2},l_{k_2}),\mathcal{S}_{n_{k_2}}}\right)^\top \\ &\quad -\sum_{\substack{\mathbf{i},\mathbf{i}'\in\{0,1\}^K\\ \|\mathbf{i}\|_1+\|\mathbf{i}'\|_1\neq 0}}\eta^{2K-\|\mathbf{i}\|_1-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}\|_1+\|\mathbf{i}'\|_1} \\ &\qquad \left(\underset{k_1\in[K]}{\circledast}\mathbf{D}_{(k_1,i_{k_1},l'_{k_1}),\mathcal{S}_{n_{k_1}}}\right)\left(\underset{k_2\in[K]}{\circledast}\mathbf{D}_{(k_2,i'_{k_2},l'_{k_2}),\mathcal{S}_{n_{k_2}}}\right)^\top. \end{aligned} \tag{113}$$

We now note that

$$\begin{aligned} \|\mathbf{A}_1 * \mathbf{A}_2\|_2 &= \|(\mathbf{A}_1\otimes\mathbf{A}_2)\mathbf{P}\|_2 \\ &\leq \|(\mathbf{A}_1\otimes\mathbf{A}_2)\|_2\|\mathbf{P}\|_2 \\ &\overset{(a)}{=} \|\mathbf{A}_1\|_2\|\mathbf{A}_2\|_2, \end{aligned} \tag{114}$$

where $\mathbf{P} \in \mathbb{R}^{p\times s}$ is a selection matrix that selects s columns of $\mathbf{A}_1\otimes\mathbf{A}_2$ and $\mathbf{p}_j = \mathbf{e}_i$ for $j \in [s], i \in [p]$. Here, (a) follows from the fact that $\|\mathbf{P}\|_2 = 1$ ($\mathbf{P}^\top\mathbf{P} = \mathbf{I}_s$). From (41), it is apparent that $\sqrt{\frac{s\varepsilon'}{r^2}} \leq 1$. Furthermore,

$$\left\|\mathbf{D}_{(k,0),\mathcal{S}_{n_k}}\right\|_2 \leq \sqrt{\frac{3}{2}}, \left\|\mathbf{D}_{(k,1,l_k),\mathcal{S}_{n_k}}\right\|_2 \leq \sqrt{\frac{s}{r^{2/K}}},\ k\in[K], \tag{115}$$

where the fist inequality in (115) follows from the RIP condition for $\left\{\mathbf{D}_{(0,k)}, k\in[K]\right\}$ and the second inequality follows from the fact that $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F$. We therefore have

$$\frac{1}{\sigma_a^2}\left\|\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}\right\|_2$$

[9] Due to the fact that $\mathcal{S}_{n_k}$'s can be multisets, $\mathbf{D}_{(k,l_k),\mathcal{S}_{n_k}}$'s can have duplicated columns.

$$
\begin{aligned}
&\overset{(b)}{\leq} 2 \sum_{\substack{\mathbf{i},\mathbf{i}'\in\{0,1\}^K \\ \|\mathbf{i}\|_1+\|\mathbf{i}'\|_1\neq 0}} \eta^{2K-\|\mathbf{i}\|_1-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}\|_1+\|\mathbf{i}'\|_1} \\
&\quad \Big\| \underset{k_1\in[K]}{\circledast} \mathbf{D}_{(k_1,i_{k_1},l_{k_1}),\mathcal{S}_{n_{k_1}}} \Big\|_2 \Big\| \underset{k_2\in[K]}{\circledast} \mathbf{D}_{(k_2,i'_{k_2},l_{k_2}),\mathcal{S}_{n_{k_2}}} \Big\|_2 \\
&\overset{(c)}{\leq} 2 \sum_{\substack{\mathbf{i}\in\{0,1\}^K \\ \|\mathbf{i}\|_1\neq 0}} \eta^{K-\|\mathbf{i}\|_1}\nu^{\|\mathbf{i}\|_1} \\
&\qquad \prod_{\substack{k_1\in[K] \\ i_{k_1}=0}} \big\|\mathbf{D}_{(k_1,0),\mathcal{S}_{n_{k_1}}}\big\|_2 \prod_{\substack{k_1\in[K] \\ i_{k_1}=1}} \big\|\mathbf{D}_{(k_1,1,l_{k_1}),\mathcal{S}_{n_{k_1}}}\big\|_2 \\
&\qquad \Big(\sum_{\mathbf{i}'\in\{0,1\}^K} \eta^{K-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}'\|_1} \\
&\qquad\quad \prod_{\substack{k_2\in[K] \\ i'_{k_2}=0}} \big\|\mathbf{D}_{(k_2,0),\mathcal{S}_{n_{k_2}}}\big\|_2 \prod_{\substack{k_2\in[K] \\ i'_{k_2}=1}} \big\|\mathbf{D}_{(k_2,1,l_{k_2}),\mathcal{S}_{n_{k_2}}}\big\|_2 \Big) \\
&+ 2 \sum_{\substack{\mathbf{i}\in\{0,1\}^K \\ \|\mathbf{i}\|_1\neq 0}} \eta^{K-\|\mathbf{i}\|_1}\nu^{\|\mathbf{i}\|_1} \\
&\qquad \prod_{\substack{k_1\in[K] \\ i_{k_1}=0}} \big\|\mathbf{D}_{(k_1,0),\mathcal{S}_{n_{k_1}}}\big\|_2 \prod_{\substack{k_1\in[K] \\ i_{k_1}=1}} \big\|\mathbf{D}_{(k_1,1,l_{k_1}),\mathcal{S}_{n_{k_1}}}\big\|_2 \\
&\Big(\sum_{\substack{\mathbf{i}'\in\{0,1\}^K \\ \|\mathbf{i}'\|_1\neq 0}} \eta^{K-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}'\|_1} \\
&\qquad\quad \prod_{\substack{k_2\in[K] \\ i'_{k_2}=0}} \big\|\mathbf{D}_{(k_2,0),\mathcal{S}_{n_{k_2}}}\big\|_2 \prod_{\substack{k_2\in[K] \\ i'_{k_2}=1}} \big\|\mathbf{D}_{(k_2,1,l_{k_2}),\mathcal{S}_{n_{k_2}}}\big\|_2 \Big) \\
&\overset{(d)}{=} 2\Big(\sum_{k_1=0}^{K-1} \binom{K}{k_1} \eta^{k_1}\nu^{K-k_1} \Big(\sqrt{\frac{3}{2}}\Big)^{k_1} \Big(\sqrt{\frac{s}{r^{2/K}}}\Big)^{K-k_1} \Big) \\
&\qquad \Big(\sum_{k_2=0}^{K} \binom{K}{k_2} \eta^{k_2}\nu^{K-k_2} \Big(\sqrt{\frac{3}{2}}\Big)^{k_2} \Big(\sqrt{\frac{s}{r^{2/K}}}\Big)^{K-k_2} \Big) \\
&+ 2\Big(\eta\sqrt{\frac{3}{2}}\Big)^{K} \Big(\sum_{k_2=0}^{K-1} \binom{K}{k_2} \eta^{k_2}\nu^{K-k_2} \Big(\sqrt{\frac{3}{2}}\Big)^{k_2} \\
&\qquad\qquad \Big(\sqrt{\frac{s}{r^{2/K}}}\Big)^{K-k_2} \Big) \\
&\overset{(e)}{\leq} 2\Big(\sum_{k_1=0}^{K-1} \binom{K}{k_1} \Big(\sqrt{\frac{3}{2}}\Big)^{k_1} \Big(\sqrt{\frac{s\varepsilon'}{r^2}}\Big)^{K-k_1} \Big) \\
&\qquad \Big(\sum_{k_2=0}^{K} \binom{K}{k_2} \Big(\sqrt{\frac{3}{2}}\Big)^{k_2} \Big(\sqrt{\frac{s\varepsilon'}{r^2}}\Big)^{K-k_2} \Big) \\
&\quad + 2\Big(\sqrt{\frac{3}{2}}\Big)^{K} \Big(\sum_{k_2=0}^{K-1} \binom{K}{k_2} \Big(\sqrt{\frac{3}{2}}\Big)^{k_2} \Big(\sqrt{\frac{s\varepsilon'}{r^2}}\Big)^{K-k_2} \Big) \\
&= 2\sqrt{\frac{s\varepsilon'}{r^2}}\Big(\sum_{k_1=0}^{K-1} \binom{K}{k_1} \Big(\sqrt{\frac{3}{2}}\Big)^{k_1} \Big(\sqrt{\frac{s\varepsilon'}{r^2}}\Big)^{K-1-k_1} \Big) \\
&\qquad \Big(\sum_{k_2=0}^{K} \binom{K}{k_2} \Big(\sqrt{\frac{3}{2}}\Big)^{k_2} \Big(\sqrt{\frac{s\varepsilon'}{r^2}}\Big)^{K-k_2} + \Big(\sqrt{\frac{3}{2}}\Big)^{K} \Big) \\
&\overset{(f)}{\leq} 2\sqrt{\frac{s\varepsilon'}{r^2}}\Big(\Big(\sqrt{\frac{3}{2}}\Big)^{K-1} \sum_{k_1=0}^{K} \binom{K}{k_1} \Big) \\
&\qquad \Big(\Big(\sqrt{\frac{3}{2}}+1\Big)^{K} + \Big(\sqrt{\frac{3}{2}}\Big)^{K} \Big) \\
&\leq 2\sqrt{\frac{s\varepsilon'}{r^2}}\Big(\Big(\sqrt{\frac{3}{2}}\Big)^{K-1} 2^K \Big) \Big(\Big(\frac{3}{2}\Big)^{K} 2^K + \Big(\frac{3}{2}\Big)^{K} \Big) \\
&\leq 3^{2K+1}\sqrt{\frac{s\varepsilon'}{r^2}}, \qquad (116)
\end{aligned}
$$

where (b) follows from triangle inequality, (c) follows from (114), (d) follows from (115), (e) and (f) follow from replacing the value for ν and the fact that $\eta < 1$ and $s\varepsilon'/r^2 < 1$ (by assumption). Denoting the smallest eigenvalue of $\mathbf{\Sigma}_{(n,l)}$ as $\lambda_{\min}(\mathbf{\Sigma}_{(n,l)})$, $\lambda_{\min}(\mathbf{\Sigma}_{(n,l)}) \geq \sigma^2$ holds; thus, we have $\|\mathbf{\Sigma}^{-1}_{(n,l)}\|_2 \leq \frac{1}{\sigma^2}$ and from [48], we get

$$
\begin{aligned}
\Big\|\mathbf{\Sigma}^{-1}_{(n,l)} - \mathbf{\Sigma}^{-1}_{(n,l')}\Big\|_2 &\leq 2\Big\|\mathbf{\Sigma}^{-1}_{(n,l)}\Big\|_2^2 \big\|\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\big\|_2 \\
&\leq \frac{2}{\sigma^4}\big\|\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\big\|_2 . \qquad (117)
\end{aligned}
$$

Now (110) can be stated as

$$
\begin{aligned}
I(\mathbf{Y};l|\mathbf{T}(\mathbf{X})) &\leq \frac{4Ns}{\sigma^4 L^2}\sum_{l,l'}\big\|\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\big\|_2^2 \\
&\leq \frac{4Ns}{\sigma^4}\big\|\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\big\|_2^2 \\
&\overset{(g)}{\leq} \frac{4Ns}{\sigma^4}(3^{4K+2})\left(\sigma_a^2\sqrt{\frac{s\varepsilon'}{r^2}}\right)^2 \\
&= 36(3^{4K})\left(\frac{\sigma_a}{\sigma}\right)^4\frac{Ns^2}{r^2}\varepsilon', \qquad (118)
\end{aligned}
$$

where (g) follow from (116). Thus, the proof is complete. ■

Proof of Lemma 5: Similar to Lemma 4, the first part of this Lemma trivially follows from Lemma 2. Also, in this case the coefficient vector is assumed to be sparse according to (24). Hence, conditioned on $\mathcal{S}_n = \mathrm{supp}(\mathbf{x}_n)$, observations $\mathbf{y}_n$'s are zero-mean independent multivariate Gaussian random vectors with covariances given by (109). Similar to Lemma 4, therefore, the conditional MI has the upper bound given in (110). We now simplify this upper bound further.

When non-zero elements of the coefficient vector are selected according to (24) and (38), we can write the dictionary $\mathbf{D}_{l,\mathcal{S}_n}$ in terms of the Kronecker product of matrices:

$$
\mathbf{D}_{l,\mathcal{S}_n} = \bigotimes_{k\in[K]} \mathbf{D}_{(k,l_k),\mathcal{S}_{n_k}}, \qquad (119)
$$

where $\mathcal{S}_{n_k} = \{j_{n_k}\}_{n_k=1}^{s_k}, j_{n_k} \in [p_k]$, for all $k \in [K]$, denotes the support of $\mathbf{x}_n$ on coordinate dictionary $\mathbf{D}_{(k,l_k)}$ and $\mathcal{S}_n$ corresponds to indexing of the elements of $(\mathcal{S}_1 \times \cdots \times \mathcal{S}_K)$. Note that $\mathbf{D}_{l,\mathcal{S}_n} \in \mathbb{R}^{(\prod_{k\in[K]} m_k)\times s}$. In contrast to coefficient model (23), in this model the $\mathcal{S}_{n_k}$'s are not multisets anymore since for each $\mathbf{D}_{(k,l_k)}, k \in [K]$, we select s_k columns at random and $\mathbf{D}_{(k,l_k),\mathcal{S}_{n_k}}$ are submatrices of $\mathbf{D}_{(k,l_k)}$. Therefore, (109) can be written as

$$
\mathbf{\Sigma}_{(n,l)} = \sigma_a^2\Big(\bigotimes_{k_1\in[K]} \mathbf{D}_{(k_1,l_{k_1}),\mathcal{S}_{n_{k_1}}}\Big)
$$

$$\left(\bigotimes_{k_2\in[K]} \mathbf{D}_{(k_2,l_{k_2}),\mathcal{S}_{n_{k_2}}} \right)^\top + \sigma^2 \mathbf{I}_s. \tag{120}$$

In order to find an upper bound for $\|\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\|_2$, notice that the expression for $\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}$ is similar to that of (113), where $\divideontimes$ is replaced by $\bigotimes$. Using the property of Kronecker product that $\|\mathbf{A}_1 \otimes \mathbf{A}_2\|_2 = \|\mathbf{A}_1\|_2 \|\mathbf{A}_2\|_2$ and the fact that

$$\left\|\mathbf{D}_{(k,0),\mathcal{S}_{n_k}}\right\|_2 \le \sqrt{\frac{3}{2}}, \left\|\mathbf{D}_{(k,1,l_k),\mathcal{S}_{n_k}}\right\|_2 \le \sqrt{\frac{s_k}{r^{2/K}}}, \forall k \in [K], \tag{121}$$

we have

$$\begin{aligned}
&\frac{1}{\sigma_a^2}\left\|\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\right\|_2 \\
&\le 2 \sum_{\substack{\mathbf{i},\mathbf{i}'\in\{0,1\}^K \\ \|\mathbf{i}\|_1+\|\mathbf{i}'\|_1\neq 0}} \eta^{2K-\|\mathbf{i}\|_1-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}\|_1+\|\mathbf{i}'\|_1} \\
&\quad \left\| \bigotimes_{k_1\in[K]} \mathbf{D}_{(k_1,i_{k_1},l_{k_1}),\mathcal{S}_{n_{k_1}}} \right\|_2 \left\| \bigotimes_{k_2\in[K]} \mathbf{D}_{(k_2,i'_{k_2},l_{k_2}),\mathcal{S}_{n_{k_2}}} \right\|_2 \\
&= 2 \sum_{\substack{\mathbf{i}\in\{0,1\}^K \\ \|\mathbf{i}\|_1\neq 0}} \eta^{K-\|\mathbf{i}\|_1}\nu^{\|\mathbf{i}\|_1} \\
&\quad \prod_{\substack{k_1\in[K] \\ i_{k_1}=0}} \left\|\mathbf{D}_{(k_1,0),\mathcal{S}_{n_{k_1}}}\right\|_2 \prod_{\substack{k_1\in[K] \\ i_{k_1}=1}} \left\|\mathbf{D}_{(k_1,1,l_{k_1}),\mathcal{S}_{n_{k_1}}}\right\|_2 \\
&\quad \Bigg(\sum_{\mathbf{i}'\in\{0,1\}^K} \eta^{K-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}'\|_1} \\
&\qquad \prod_{\substack{k_2\in[K] \\ i'_{k_2}=0}} \left\|\mathbf{D}_{(k_2,0),\mathcal{S}_{n_{k_2}}}\right\|_2 \prod_{\substack{k_2\in[K] \\ i'_{k_2}=1}} \left\|\mathbf{D}_{(k_2,1,l_{k_2}),\mathcal{S}_{n_{k_2}}}\right\|_2 \Bigg) \\
&+ 2\Bigg(\eta^K \prod_{k_1\in[K]} \left\|\mathbf{D}_{(k_1,0),\mathcal{S}_{n_{k_1}}}\right\|_2\Bigg)\Bigg(\sum_{\substack{\mathbf{i}'\in\{0,1\}^K \\ \|\mathbf{i}'\|_1\neq 0}} \eta^{K-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}'\|_1} \\
&\qquad \prod_{\substack{k_2\in[K] \\ i'_{k_2}=0}} \left\|\mathbf{D}_{(k_2,0),\mathcal{S}_{n_{k_2}}}\right\|_2 \prod_{\substack{k_2\in[K] \\ i'_{k_2}=1}} \left\|\mathbf{D}_{(k_2,1,l_{k_2}),\mathcal{S}_{n_{k_2}}}\right\|_2 \Bigg) \\
&\overset{(a)}{\le} 2\sqrt{s}\Bigg[\Bigg(\sum_{k_1=0}^{K-1} \binom{K}{k_1} \eta^{k_1}\nu^{K-k_1} \left(\sqrt{\frac{3}{2}}\right)^{k_1} \left(\sqrt{\frac{1}{r^{2/K}}}\right)^{K-k_1} \Bigg) \\
&\quad \Bigg(\sum_{k_2=0}^{K} \binom{K}{k_2} \left(\eta\sqrt{\frac{3}{2}}\right)^{k_2} \Bigg) + \left(\eta\sqrt{\frac{3}{2}}\right)^K \\
&\quad \Bigg(\sum_{k_2=0}^{K-1} \binom{K}{k_2} \eta^{k_2}\nu^{(K-k_2)} \left(\sqrt{\frac{3}{2}}\right)^{k_2} \left(\sqrt{\frac{1}{r^{2/K}}}\right)^{K-k_2} \Bigg)\Bigg] \\
&\overset{(b)}{\le} 2\sqrt{\frac{s\varepsilon'}{r^2}}\Bigg(\sum_{k_1=0}^{K-1} \binom{K}{k_1} \left(\sqrt{\frac{3}{2}}\right)^{k_1} \Bigg) \\
&\quad \Bigg(\Bigg(\sum_{k_2=0}^{K} \binom{K}{k_2} \left(\sqrt{\frac{3}{2}}\right)^{k_2} \Bigg) + \left(\sqrt{\frac{3}{2}}\right)^K \Bigg) \\
&\overset{(c)}{\le} 3^{2K+1}\sqrt{\frac{s\varepsilon'}{r^2}},
\end{aligned} \tag{122}$$

where (a) follows from (121), (b) follows from replacing the value for ν and the fact that $\eta < 1$, $\varepsilon'/r^2 < 1$ (by assumption), and (c) follows from similar arguments in (116). The rest of the proof follows the same arguments as in Lemma 4 and (118) holds in this case as well. ∎

Proof of Theorem 4: Any dictionary $\mathbf{D} \in \mathcal{X}(\mathbf{I}_p, r)$ can be written as

$$\begin{aligned}
\mathbf{D} &= \mathbf{A} \otimes \mathbf{B} \\
&= (\mathbf{I}_{p_1} + \mathbf{\Delta}_1) \otimes (\mathbf{I}_{p_2} + \mathbf{\Delta}_2),
\end{aligned} \tag{123}$$

We have to ensure that $\|\mathbf{D} - \mathbf{I}_p\|_F \le r$. We have

$$\begin{aligned}
&\|\mathbf{D} - \mathbf{I}_p\|_F \\
&\quad = \|\mathbf{I}_{p_1} \otimes \mathbf{\Delta}_2 + \mathbf{\Delta}_1 \otimes \mathbf{I}_{p_2} + \mathbf{\Delta}_1 \otimes \mathbf{\Delta}_2\|_F \\
&\quad \le \|\mathbf{I}_{p_1} \otimes \mathbf{\Delta}_2\|_F + \|\mathbf{\Delta}_1 \otimes \mathbf{I}_{p_2}\|_F + \|\mathbf{\Delta}_1 \otimes \mathbf{\Delta}_2\|_F \\
&\quad = \|\mathbf{I}_{p_1}\|_F\|\mathbf{\Delta}_2\|_F + \|\mathbf{\Delta}_1\|_F\|\mathbf{I}_{p_2}\|_F + \|\mathbf{\Delta}_1\|_F\|\mathbf{\Delta}_2\|_F \\
&\quad \le r_2\sqrt{p_1} + r_1\sqrt{p_2} + r_1 r_2 \\
&\quad \overset{(a)}{\le} r,
\end{aligned} \tag{124}$$

where (a) follows from (51). Therefore, we have

$$\begin{aligned}
\mathbf{D} \in \Big\{ &\mathbf{A} \otimes \mathbf{B} = (\mathbf{I}_{p_1} + \mathbf{\Delta}_1) \otimes (\mathbf{I}_{p_2} + \mathbf{\Delta}_2) \big|\ \|\mathbf{\Delta}_1\|_F \le r_1, \\
&\|\mathbf{\Delta}_2\|_F \le r_2,\ r_2\sqrt{p_1} + r_1\sqrt{p_2} + r_1 r_2 \le r, \\
&\|\mathbf{a}_{l_1}\|_2 = 1, l_1 \in [p_1],\ \|\mathbf{b}_{l_2}\|_2 = 1, l_2 \in [p_2] \Big\}.
\end{aligned} \tag{125}$$

In this case, the new observation vectors $\mathbf{y}'_{(n,j)}$ can be written as

$$\mathbf{y}'_{(n,j)} = \mathbf{A}\mathbf{x}'_{(n,j)} + \mathbf{A}_p\mathbf{x}_n,\ j \in [p_2],\ n \in [N], \tag{126}$$

where $\mathbf{A}_p \triangleq (\mathbf{A} \otimes \mathbf{\Delta}_2)^{\mathcal{T}_n}$ denotes the matrix consisting of the rows of $(\mathbf{A} \otimes \mathbf{\Delta}_2)$ with indices $\mathcal{T}_n \triangleq ip_2 + j$, where $i = \{0\} \cup [p_1 - 1]$ and $j = \big((n-1) \bmod p_2\big) + 1$.

Similarly, for $\mathbf{y}''_{(n,j)}$ we have

$$\mathbf{y}''_{(n,j)} = \mathbf{B}\mathbf{x}''_{(n,j)} + \mathbf{B}_p\mathbf{x}_n,\ j \in [p_1],\ n \in [N], \tag{127}$$

where $\mathbf{B}_p \triangleq (\mathbf{\Delta}_1 \otimes \mathbf{B})^{\mathcal{I}_n}$ denotes the matrix consisting of the rows of $(\mathbf{\Delta}_1 \otimes \mathbf{B})$ with indices $\mathcal{I}_n \triangleq jp_2 + i$, where $i = \{0\} \cup [p_2 - 1]$ and $j = (n-1) \bmod p_1$. Given the fact that $\mathbf{x}_n \in \{-1, 0, 1\}^p$, $\sigma_a^2 = 1$ and $\|\mathbf{x}_n\|_2^2 = s$, after division of the coefficient vector according to (55) and (59), we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}_n}\left\{x_{n,l}^2\right\} &= \mathbb{E}_{\mathbf{x}'_{(n,j_1)}}\left\{{x'}^2_{(n,j_1),l_1}\right\} = \mathbb{E}_{\mathbf{x}''_{(n,j_2)}}\left\{{x''}^2_{(n,j_2),l_2}\right\} \\
&= \frac{s}{p},
\end{aligned} \tag{128}$$

for any $n \in [N], j_1 \in [p_2], j_2 \in [p_1], l \in [p], l_1 \in [p_1]$, and $l_2 \in [p_2]$. The SNR is

$$\mathrm{SNR} = \frac{\mathbb{E}_{\mathbf{x}}\left\{\|\mathbf{x}\|_2^2\right\}}{\mathbb{E}_{\boldsymbol{\eta}}\left\{\|\boldsymbol{\eta}\|_2^2\right\}} = \frac{s}{m\sigma^2}. \tag{129}$$

We are interested in upper bounding $\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\right\|_F^2\right\}$.

For this purpose we first upper bound $\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{A}}(\mathbf{Y})-\mathbf{A}\right\|_F^2\right\}$ and $\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{B}}(\mathbf{Y})-\mathbf{B}\right\|_F^2\right\}$. We can split these MSEs into the sum of column-wise MSEs:

$$\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{A}}(\mathbf{Y})-\mathbf{A}\right\|_F^2\right\}=\sum_{l=1}^{p_1}\mathbb{E}_{\mathbf{Y}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\right\}. \quad (130)$$

By construction:

$$\begin{aligned}\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2 &\le 2\left(\|\widehat{\mathbf{a}}_l(\mathbf{Y})\|_2^2+\|\mathbf{a}_l\|_2^2\right)\\ &\overset{(b)}{\le} 4, \end{aligned}\quad (131)$$

where (b) follows from the projection step in (62). We define the event $\mathcal{C}$ to be

$$\mathcal{C}\triangleq\bigcap_{\substack{n\in[N]\\ l\in[p]}}\{|\eta_{n,l}|\le 0.4\}. \quad (132)$$

In order to find the setting under which $\mathbb{P}\left\{\widehat{\mathbf{X}}=\mathbf{X}|\mathcal{C}\right\}=1$, i.e., when recovery of the coefficient vectors is successful, we observe the original observations and coefficient vectors satisfy:

$$y_{n,l}-x_{n,l}=\left(\mathbf{I}_{p_1}\otimes\mathbf{\Delta}_2+\mathbf{\Delta}_1\otimes\mathbf{I}_{p_2}+\mathbf{\Delta}_1\otimes\mathbf{\Delta}_2\right)^l\mathbf{x}_n+\eta_{n,l} \quad (133)$$

and

$$\begin{aligned}&\left|\left(\mathbf{I}_{p_1}\otimes\mathbf{\Delta}_2+\mathbf{\Delta}_1\otimes\mathbf{I}_{p_2}+\mathbf{\Delta}_1\otimes\mathbf{\Delta}_2\right)^l\mathbf{x}_n+\eta_{n,l}\right|\\ &\le\left\|\left(\mathbf{I}_{p_1}\otimes\mathbf{\Delta}_2+\mathbf{\Delta}_1\otimes\mathbf{I}_{p_2}+\mathbf{\Delta}_1\otimes\mathbf{\Delta}_2\right)^l\right\|_2\|\mathbf{x}_n\|_2+|\eta_{n,l}|\\ &\le\left(\|\mathbf{\Delta}_1\|_F+\|\mathbf{\Delta}_2\|_F+\|\mathbf{\Delta}_1\|_F\|\mathbf{\Delta}_2\|_F\right)\|\mathbf{x}_n\|_2+|\eta_{n,l}|\\ &\le(r_1+r_2+r_1r_2)\sqrt{s}+|\eta_{n,l}|. \end{aligned}\quad (134)$$

By using the assumption $(r_1+r_2+r_1r_2)\sqrt{s}\le 0.1$ and conditioned on the event $\mathcal{C}$, $|\eta_{n,l}|\le 0.4$, we have that for every $n\in[N]$ and $l\in[p]$:

$$\begin{cases} y_{n,l}>0.5 & \text{if } x_{n,l}=1,\\ -0.5<y_{n,l}<0.5 & \text{if } x_{n,l}=0,\\ y_{n,l}<-0.5 & \text{if } x_{n,l}=-1, \end{cases}\quad (135)$$

thus, ensuring correct recovery of coefficients ($\widehat{\mathbf{X}}=\mathbf{X}$) using the thresholding technique (53) when conditioned on $\mathcal{C}$. Using standard tail bounds for Gaussian random variables [34, (92)], [49, Proposition 7.5] and taking a union bound over all pN i.i.d. variables $\{\eta_{n,l}\}, n\in[N], l\in[p]$, we have

$$\mathbb{P}\{\mathcal{C}^c\}\le\exp\left(-\frac{0.08pN}{\sigma^2}\right). \quad (136)$$

To find an upper bound for $\mathbb{E}_{\mathbf{Y}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\right\}$, we can write it as

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\right\}&=\mathbb{E}_{\mathbf{Y},\mathbf{N}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2|\mathcal{C}\right\}\mathbb{P}(\mathcal{C})\\ &\quad+\mathbb{E}_{\mathbf{Y},\mathbf{N}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2|\mathcal{C}^c\right\}\mathbb{P}(\mathcal{C}^c)\\ &\overset{(c)}{\le}\mathbb{E}_{\mathbf{Y},\mathbf{N}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2|\mathcal{C}\right\}+4\exp\left(-\frac{0.08pN}{\sigma^2}\right), \end{aligned}\quad (137)$$

where (c) follows from (131) and (136). To bound $\mathbb{E}_{\mathbf{Y},\mathbf{N}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2|\mathcal{C}\right\}$, we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{Y},\mathbf{N}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2|\mathcal{C}\right\}\\ &=\mathbb{E}_{\mathbf{Y},\mathbf{N}}\left\{\|P_{\mathcal{B}_1}(\widetilde{\mathbf{a}}_l(\mathbf{Y}))-\mathbf{a}_l\|_2^2|\mathcal{C}\right\}\\ &\overset{(d)}{\le}\mathbb{E}_{\mathbf{Y},\mathbf{N}}\left\{\|\widetilde{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2|\mathcal{C}\right\}\\ &\overset{(e)}{=}\mathbb{E}_{\mathbf{Y},\mathbf{N}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}\widehat{x'}_{(n,j),l}\mathbf{y}'_{(n,j)}-\mathbf{a}_l\right\|_2^2\Big|\mathcal{C}\right\}\\ &\overset{(f)}{=}\mathbb{E}_{\mathbf{Y},\mathbf{X},\mathbf{N}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\mathbf{y}'_{(n,j)}-\mathbf{a}_l\right\|_2^2\Big|\mathcal{C}\right\}\\ &\overset{(g)}{=}\mathbb{E}_{\mathbf{X},\mathbf{N}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\big(\mathbf{A}\mathbf{x}'_{(n,j)}+\mathbf{A}_p\mathbf{x}_n\right.\right.\\ &\qquad\left.\left.+\boldsymbol{\eta}'_{(n,j)}\big)-\mathbf{a}_l\right\|_2^2\Big|\mathcal{C}\right\}\\ &\overset{(h)}{\le}2\mathbb{E}_{\mathbf{X},\mathbf{N}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\boldsymbol{\eta}'_{(n,j)}\right\|_2^2\Big|\mathcal{C}\right\}\\ &\quad+4\mathbb{E}_{\mathbf{X},\mathbf{N}}\left\{\left\|\mathbf{a}_l-\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\sum_{t=1}^{p_1}\mathbf{a}_t x'_{(n,j),t}\right\|_2^2\Big|\mathcal{C}\right\}\\ &\quad+4\mathbb{E}_{\mathbf{X},\mathbf{N}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\sum_{t=1}^{p}\mathbf{a}_{p,t}x_{n,t}\right\|_2^2\Big|\mathcal{C}\right\}, \end{aligned}\quad (138)$$

where (d) follows from the fact that $\|\mathbf{a}_l\|_2=1$, (e) follows from (57), (f) follows from the fact that conditioned on the event $\mathcal{C}$, $\widehat{\mathbf{X}}=\mathbf{X}$, (g) follows from (126) and (h) follows from the fact that $\|\mathbf{x}_1+\mathbf{x}_2\|_2^2\le 2(\|\mathbf{x}_1\|_2^2+\|\mathbf{x}_2\|_2^2)$. We bound the three terms in (138) separately. Defining $\nu\triangleq\mathcal{Q}(-0.4/\sigma)-\mathcal{Q}(0.4/\sigma)$, where $\mathcal{Q}(x)\triangleq\int_{z=x}^{\infty}\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})dz$, we can bound the noise variance conditioned on $\mathcal{C}$, $\sigma^2_{\eta_{n,t}}$, by [34]

$$\sigma^2_{\eta_{n,t}}\le\frac{\sigma^2}{\nu}. \quad (139)$$

The first expectation in (138) can be bounded by

$$\begin{aligned}&\mathbb{E}_{\mathbf{X},\mathbf{N}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\boldsymbol{\eta}'_{(n,j)}\right\|_2^2\Big|\mathcal{C}\right\}\\ &=\left(\frac{p_1}{Ns}\right)^2\sum_{n,n'=1}^{N}\sum_{j,j'=1}^{p_2}\mathbb{E}_{\mathbf{X},\mathbf{N}}\left\{x'_{(n,j),l}x'_{(n',j'),l}\right.\\ &\qquad\left.\boldsymbol{\eta}'^{\top}_{(n',j')}\boldsymbol{\eta}'_{(n,j)}|\mathcal{C}\right\}\\ &=\left(\frac{p_1}{Ns}\right)^2\sum_{n=1}^{N}\sum_{j=1}^{p_2}\sum_{t=1}^{m_1}\mathbb{E}_{\mathbf{X},\mathbf{N}}\left\{x'^2_{(n,j),l}|\mathcal{C}\right\}\mathbb{E}_{\mathbf{X},\mathbf{N}}\left\{\eta'^2_{(n,j),t}|\mathcal{C}\right\}\\ &\overset{(i)}{=}\left(\frac{p_1}{Ns}\right)^2Np_2\mathbb{E}_{\mathbf{X}}\left\{x'^2_{(n,j),l}\right\}\mathbb{E}_{\mathbf{N}}\left\{\eta'^2_{(n,j),t}|\mathcal{C}\right\} \end{aligned}$$

$$\overset{(j)}{\leq} \left(\frac{p_1}{Ns}\right)^2 Np_2 \left(\frac{s}{p}\right)\left(\frac{m_1\sigma^2}{\nu}\right)$$
$$\overset{(k)}{\leq} \frac{2m_1p_1\sigma^2}{Ns}, \tag{140}$$

where (i) follows from the fact that $\mathbf{x}'_{(n,j)}$ is independent of the event $\mathcal{C}$, (j) follows from (128) and (139), and (k) follows from the fact that $\nu \geq 0.5$ under the assumption that $\sigma \leq 0.4$ [34].

To bound the second expectation in (138), we use similar arguments as in Jung et al. [34]. We can write

$$\mathbb{E}_{\mathbf{X}}\{x'_{(n,j),l}x'_{(n,j),t}x'_{(n',j'),l}x'_{(n',j'),t'}\} = \begin{cases} (\frac{s}{p})^2 & \text{if } (n,j)=(n',j') \text{ and } t=t'\neq l, \\ (\frac{s}{p})^2 & \text{if } (n,j)\neq(n',j') \text{ and } t=t'=l, \\ \frac{s}{p} & \text{if } (n,j)=(n',j') \text{ and } t=t'=l, \\ 0 & \text{otherwise,} \end{cases} \tag{141}$$

and we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X},\mathbf{N}}\left\{\left\|\mathbf{a}_l - \frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\sum_{t=1}^{p_1}\mathbf{a}_t x'_{(n,j),t}\right\|_2^2\middle|\mathcal{C}\right\} \\
&\leq \mathbf{a}_l^\top\mathbf{a}_l - \frac{2p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}\sum_{t=1}^{p_1}\mathbf{a}_l^\top\mathbf{a}_t\mathbb{E}_{\mathbf{X}}\left\{x'_{(n,j),l}x'_{(n,j),t}\right\} \\
&\quad + \left(\frac{p_1}{Ns}\right)^2\sum_{n,n'=1}^{N}\sum_{j,j'=1}^{p_2}\sum_{t,t'=1}^{p_1}\mathbf{a}_{t'}^\top\mathbf{a}_t \\
&\qquad\qquad \mathbb{E}_{\mathbf{X}}\left\{x'_{(n',j'),l}x'_{(n',j'),t'}x'_{(n,j),l}x'_{(n,j),t}\right\} \\
&= 1 - \left(\frac{2p_1}{Ns}\right)(p_2N)\left(\frac{s}{p}\right) + \left(\frac{p_1}{Ns}\right)^2(p_2N) \\
&\qquad \left(\frac{s}{p} + (p_1-1)\left(\frac{s}{p}\right)^2 + (p_2N-1)\left(\frac{s}{p}\right)^2\right) \\
&= \frac{p_1}{N}\left(\frac{1}{s} + \frac{1}{p_2} - \frac{2}{p}\right) \\
&\leq \frac{2p_1}{N}.
\end{aligned} \tag{142}$$

To upper bound the third expectation in (138), we need to bound the ℓ_2 norm of columns of $\mathbf{A}_p$. We have

$$\begin{aligned}
\forall t \in [p] : \|\mathbf{a}_{p,t}\|_2^2 &\overset{(l)}{\leq} \|(\mathbf{A}\otimes\mathbf{\Delta}_2)_t\|_2^2 \\
&\leq \|\mathbf{a}_l\|_2^2\|\mathbf{\Delta}_2\|_F^2 \\
&= r_2^2,
\end{aligned} \tag{143}$$

where $(\mathbf{A}\otimes\mathbf{\Delta}_2)_t$ denotes the t-th column of $(\mathbf{A}\otimes\mathbf{\Delta}_2)$ and (l) follows from the fact that $\mathbf{A}_p$ is a submatrix of $(\mathbf{A}\otimes\mathbf{\Delta}_2)$. Moreover, similar to the expectation in (141), we have

$$\mathbb{E}_{\mathbf{X}}\{x'_{(n,j),l}x'_{(n',j'),l}x_{n,t}x_{n',t'}\} = \begin{cases} (\frac{s}{p})^2 & \text{if } (n,j)=(n',j') \text{ and } t=t'\neq l', \\ (\frac{s}{p})^2 & \text{if } (n,j)\neq(n',j') \text{ and } t=t'=l', \\ \frac{s}{p} & \text{if } (n,j)=(n',j') \text{ and } t=t'=l', \\ 0 & \text{Otherwise,} \end{cases} \tag{144}$$

where l' denotes the index of the element of $\mathbf{x}_n$ corresponding to $x'_{(n,j),l}$. Then, the expectation can be bounded by

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X},\mathbf{N}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\sum_{t=1}^{p}\mathbf{a}_{p,t}x_{n,t}\right\|_2^2\middle|\mathcal{C}\right\} \\
&= \left(\frac{p_1}{Ns}\right)^2\sum_{n,n'=1}^{N}\sum_{j,j'=1}^{p_2}\sum_{t,t'=1}^{p}\mathbf{a}_{p,t'}^\top\mathbf{a}_{p,t} \\
&\qquad\qquad \mathbb{E}_{\mathbf{X}}\left\{x'_{(n,j),l}x'_{(n',j'),l}x_{n,t}x_{n',t'}\right\} \\
&\overset{(m)}{\leq} r_2^2\left(\frac{p_1}{Ns}\right)^2 Np_2\left(\frac{s}{p} + (p-1)\left(\frac{s}{p}\right)^2\right. \\
&\qquad\qquad \left. + (Np_2-1)\left(\frac{s}{p}\right)^2\right) \\
&\leq r_2^2\left(\frac{p_1}{Ns} + \frac{p_1}{N} + 1\right) \\
&\overset{(n)}{\leq} \frac{p_1}{N},
\end{aligned} \tag{145}$$

where (m) follows from (143) and (n) follows from the assumption in (51). Summing up (140), (142), and (145), we have

$$\mathbb{E}_{\mathbf{Y}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y}) - \mathbf{a}_l\|_2^2\right\} \leq \frac{4p_1}{N}\left(\frac{m_1\sigma^2}{s} + 3\right) + 4\exp\left(-\frac{0.08pN}{\sigma^2}\right). \tag{146}$$

Summing up the MSE for all columns, we obtain:

$$\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{A}}(\mathbf{Y}) - \mathbf{A}\right\|_F^2\right\} \leq \frac{4p_1^2}{N}\left(\frac{m_1\sigma^2}{s} + 3\right) + 4p_1\exp\left(-\frac{0.08pN}{\sigma^2}\right). \tag{147}$$

We can follow similar steps to get

$$\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{B}}(\mathbf{Y}) - \mathbf{B}\right\|_F^2\right\} \leq \frac{4p_2^2}{N}\left(\frac{m_2\sigma^2}{s} + 3\right) + 4p_2\exp\left(-\frac{0.08pN}{\sigma^2}\right). \tag{148}$$

From (147) and (148), we get

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\right\|_F^2\right\} \\
&= \mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{A}}(\mathbf{Y})\otimes\widehat{\mathbf{B}}(\mathbf{Y}) - \mathbf{A}\otimes\mathbf{B}\right\|_F^2\right\} \\
&= \mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{A}}(\mathbf{Y}) - \mathbf{A})\otimes\widehat{\mathbf{B}}(\mathbf{Y}) + \mathbf{A}\otimes(\widehat{\mathbf{B}}(\mathbf{Y}) - \mathbf{B})\right\|_F^2\right\} \\
&\leq 2\Bigg(\mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{A}}(\mathbf{Y}) - \mathbf{A})\otimes\widehat{\mathbf{B}}(\mathbf{Y})\right\|_F^2\right\} \\
&\qquad + \mathbb{E}_{\mathbf{Y}}\left\{\left\|\mathbf{A}\otimes(\widehat{\mathbf{B}}(\mathbf{Y}) - \mathbf{B})\right\|_F^2\right\}\Bigg) \\
&\leq 2\Bigg(\mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{A}}(\mathbf{Y}) - \mathbf{A})\right\|_F^2\right\}\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{B}}(\mathbf{Y})\right\|_F^2\right\} \\
&\qquad + \|\mathbf{A}\|_F^2\,\mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{B}}(\mathbf{Y}) - \mathbf{B})\right\|_F^2\right\}\Bigg) \\
&\leq 2\Bigg(p_2\mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{A}}(\mathbf{Y}) - \mathbf{A})\right\|_F^2\right\}
\end{aligned}$$

$$+ p_1 \mathbb{E}_{\mathbf{Y}} \left\{ \left\| (\widehat{\mathbf{B}}(\mathbf{Y}) - \mathbf{B}) \right\|_F^2 \right\} \Bigg)$$
$$\leq \frac{8p}{N} \left(\frac{\sigma^2}{s} \sum_{k=1}^{2} m_k p_k + 3 \sum_{k=1}^{2} p_k \right) + 8p \exp\left(-\frac{0.08pN}{\sigma^2} \right)$$
$$\overset{(o)}{=} \frac{8p}{N} \left(\frac{\sum_{k=1}^{2} m_k p_k}{m \, \mathrm{SNR}} + 3 \sum_{k=1}^{2} p_k \right) + 8p \exp\left(-\frac{0.08pN}{\sigma^2} \right), \tag{149}$$

where (o) follows from (129). ■

REFERENCES

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

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

[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. [Online]. Available: https://dx.doi.org/10.1109/TSP.2006.881199

[4] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariance sparse coding for audio classification," in *Proc. 23rd Conf. Uncertainty in Artificial Intelligence*, July 2007, pp. 149–158. [Online]. Available: http://dl.acm.org/citation.cfm?id=3020488.3020507

[5] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Machine learning*. ACM, 2007, pp. 759–766. [Online]. Available: https://dx.doi.org/10.1145/1273496.1273592

[6] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Analys. and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, April 2012. [Online]. Available: https://dx.doi.org/10.1109/TPAMI.2011.156

[7] 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: https://dx.doi.org/10.1162/089976603762552951

[8] Z. Zhang and S. Aeron, "Denoising and completion of 3D data via multidimensional dictionary learning," in *Proc. 25th Int. Joint Conf. Artificial Intelligence (IJCAI)*, July 2016, pp. 2371–2377. [Online]. Available: https://www.ijcai.org/Proceedings/16/Papers/338.pdf

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

[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: https://dx.doi.org/10.1109/ICDSP.2013.6622725

[11] 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. (ICASSP)*, May 2014, pp. 3963–3967. [Online]. Available: https://dx.doi.org/10.1109/ICASSP.2014.6854345

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

[13] M. Ghassemi, Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "STARK: Structured dictionary learning through rank-one tensor recovery," in *Proc. IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, December 2017.

[14] 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)*, June 2014, pp. 2949–2956. [Online]. Available: https://dx.doi.org/10.1109/CVPR.2014.377

[15] S. Soltani, M. E. Kilmer, and P. C. Hansen, "A tensor-based dictionary learning approach to tomographic image reconstruction," *BIT Numerical Mathematics*, pp. 1–30, 2015. [Online]. Available: https://dx.doi.org/10.1007/s10543-016-0607-z

[16] 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)*, November 2012, pp. 493–496. [Online]. Available: https://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6460179

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

[18] 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, December 1970. [Online]. Available: https://www.psychology.uwo.ca/faculty/harshman/wpppfac0.pdf

[19] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Analy. and Applicat.*, vol. 21, no. 4, pp. 1253–1278, 2000. [Online]. Available: https://dx.doi.org/10.1137/S0895479896305696

[20] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover, "Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging," *SIAM J. Matrix Anal. and Applicat.*, vol. 34, no. 1, pp. 148–172, 2013. [Online]. Available: https://dx.doi.org/10.1137/110837711

[21] Y. Rivenson and A. Stern, "Compressed imaging with a separable sensing operator," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 449–452, June 2009. [Online]. Available: https://dx.doi.org/10.1109/LSP.2009.2017817

[22] ——, "An efficient method for multi-dimensional compressive imaging," in *Frontiers in Optics 2009/Laser Science XXV/Fall 2009 OSA Optics & Photonics Technical Diges*. Optical Society of America, 2009, p. CTuA4. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=COSI-2009-CTuA4

[23] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 494–504, Febuary 2012. [Online]. Available: https://dx.doi.org/10.1109/TIP.2011.2165289

[24] A. B. Tsybakov, *Introduction to nonparametric estimation*. New York, NJ USA: Springer Series in Statistics, Springer, 2009.

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

[26] 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 Applicat.*, vol. 416, no. 1, pp. 48–67, July 2006. [Online]. Available: https://dx.doi.org/10.1016/j.laa.2005.06.035

[27] 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.

[28] A. Agarwal, A. Anandkumar, and P. Netrapalli, "A clustering approach to learn sparsely-used overcomplete dictionaries," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 575–592, January 2017. [Online]. Available: https://dx.doi.org/10.1109/TIT.2016.2614684

[29] 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. [Online]. Available: https://www.jmlr.org/proceedings/papers/v35/arora14.pdf

[30] 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, November 2014. [Online]. Available: https://dx.doi.org/10.1016/j.acha.2014.01.005

[31] ——, "Local identification of overcomplete dictionaries," *J. Machine Learning Research*, vol. 16, pp. 1211–1242, June 2015. [Online]. Available: https://jmlr.org/papers/v16/schnass15a.html

[32] 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: https://dx.doi.org/10.1109/TIT.2015.2472522

[33] A. Jung, Y. C. Eldar, and N. Görtz, "Performance limits of dictionary learning for sparse coding," in *Proc. IEEE 22nd European Signal Process. Conf. (EUSIPCO)*, September 2014, pp. 765–769. [Online]. Available: https://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6952232

[34] ——, "On the minimax risk of dictionary learning," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1501–1515, March 2015. [Online]. Available: https://dx.doi.org/10.1109/TIT.2016.2517006
[35] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*. Cambridge University Press, 1991.
[36] A. Smilde, R. Bro, and P. Geladi, *Multi-way analysis: Applications in the chemical sciences*. John Wiley & Sons, 2005.
[37] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009. [Online]. Available: https://dx.doi.org/10.1137/07070111X
[38] 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: https://dx.doi.org/10.1162/NECO_a_00385
[39] C. F. Van Loan, "The ubiquitous Kronecker product," *J. Computational and Appl. Mathematics*, vol. 123, no. 1, pp. 85–100, November 2000. [Online]. Available: https://dx.doi.org/10.1016/S0377-0427(00)00393-9
[40] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, December 2009. [Online]. Available: https://dx.doi.org/10.1109/TIT.2009.2032816
[41] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE trans. Inf. theory*, vol. 51, no. 12, pp. 4203–4215, November 2005. [Online]. Available: https://dx.doi.org/10.1109/TIT.2005.858979
[42] 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, June 2015. [Online]. Available: https://dx.doi.org/10.1109/TIT.2015.2424238
[43] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*. New York, NY USA: Cambridge University Press, 2009.
[44] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. John Wiley & Sons, 2012.
[45] J.-L. Durrieu, J. Thiran, F. Kelly *et al.*, "Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, March 2012, pp. 4833–4836. [Online]. Available: https://dx.doi.org/10.1109/ICASSP.2012.6289001
[46] J. von Neumann, "Some matrix inequalities and metrization of matrix space," *Tomsk Univ. Rev.*, vol. 1, no. 11, pp. 286–300, 1937, Reprinted in Collected Works (Pergamon Press, 1962), iv, 205 –219.
[47] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *Proc. 2010 IEEE Int. Symp. Inf. Theory*. IEEE, July 2010, pp. 1373–1377. [Online]. Available: https://dx.doi.org/10.1109/ISIT.2010.5513573
[48] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.
[49] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Springer, 2013, vol. 1, no. 3.

Zahra Shakeri is pursuing a Ph.D. degree at Rutgers University, NJ, USA. She is a member of the INSPIRE laboratory. She received her M.Sc. degree in Electrical and Computer Engineering from Rutgers University, NJ, USA, in 2016 and her B.Sc. degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran, in 2013. Her research interests are in the areas of machine learning, statistical signal processing, and multidimensional data processing.

Waheed U. Bajwa received BE (with Honors) degree in electrical engineering from the National University of Sciences and Technology, Pakistan in 2001, and MS and PhD degrees in electrical engineering from the University of Wisconsin-Madison in 2005 and 2009, respectively. He was a Postdoctoral Research Associate in the Program in Applied and Computational Mathematics at Princeton University from 2009 to 2010, and a Research Scientist in the Department of Electrical and Computer Engineering at Duke University from 2010 to 2011. He is currently an Associate Professor in the Department of Electrical and Computer Engineering at Rutgers University. His research interests include statistical signal processing, high-dimensional statistics, machine learning, networked systems, and inverse problems.

Dr. Bajwa has received a number of awards in his career including the Best in Academics Gold Medal and Presidents Gold Medal in Electrical Engineering from the National University of Sciences and Technology (2001), the Morgridge Distinguished Graduate Fellowship from the University of Wisconsin-Madison (2003), the Army Research Office Young Investigator Award (2014), the National Science Foundation CAREER Award (2015), Rutgers University's Presidential Merit Award (2016), Rutgers Engineering Governing Council ECE Professor of the Year Award (2016, 2017), and Rutgers University's Presidential Fellowship for Teaching Excellence (2017). He is a co-investigator on the work that received the Cancer Institute of New Jersey's Gallo Award for Scientific Excellence in 2017, a co-author on papers that received Best Student Paper Awards at IEEE IVMSP 2016 and IEEE CAMSAP 2017 workshops, and a Member of the Class of 2015 National Academy of Engineering Frontiers of Engineering Education Symposium. He served as an Associate Editor of the IEEE Signal Processing Letters (2014 2017), co-guest edited a special issue of Elsevier Physical Communication Journal on "Compressive Sensing in Communications" (2012), co-chaired CPSWeek 2013 Workshop on Signal Processing Advances in Sensor Networks and IEEE GlobalSIP 2013 Symposium on New Sensing and Statistical Inference Methods, and served as the Publicity and Publications Chair of IEEE CAMSAP 2015 and General Chair of the 2017 DIMACS Workshop on Distributed Optimization, Information Processing, and Learning. He is currently Technical Co-Chair of the IEEE SPAWC 2018 Workshop and serves on the MLSP, SAM, and SPCOM Technical Committees of the IEEE Signal Processing Society.

Anand D. Sarwate (S'99–M'09–SM'14) received the B.S. degrees in electrical engineering and computer science and mathematics from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from the Department of Electrical Engineering and Computer Sciences (EECS), University of California, Berkeley (U.C. Berkeley), Berkeley, CA, USA.

He is a currently an Assistant Professor with the Department of Electrical and Computer Engineering, The State University of New Jersey, New Brunswick, NJ, USA, since January 2014. He was previously a Research Assistant Professor from 2011 to 2013 with the Toyota Technological Institute at Chicago; prior to this, he was a Postdoctoral Researcher from 2008 to 2011 with the University of California, San Diego, CA.

His research interests include information theory, machine learning, signal processing, optimization, and privacy and security. Dr. Sarwate received the NSF CAREER award in 2015, and the Samuel Silver Memorial Scholarship Award and the Demetri Angelakos Memorial Award from the EECS Department at U.C. Berkeley. He was awarded the National Defense Science and Engineering Graduate Fellowship from 2002 to 2005. He is a member of Phi Beta Kappa and Eta Kappa Nu.