Identifiability of Kronecker-Structured Dictionaries for Tensor Data

Zahra Shakeri, *Student Member, IEEE*, Anand D. Sarwate, *Senior Member, IEEE*, and Waheed U. Bajwa, *Senior Member, IEEE*

***Abstract*—This paper derives sufficient conditions for local recovery of coordinate dictionaries comprising a Kronecker-structured dictionary that is used for representing Kth-order tensor data. Tensor observations are assumed to be generated from a Kronecker-structured dictionary multiplied by sparse coefficient tensors that follow the separable sparsity model. This paper provides sufficient conditions on the underlying coordinate dictionaries, coefficient and noise distributions, and number of samples that guarantee recovery of the individual coordinate dictionaries up to a specified error, as a local minimum of the objective function, with high probability. In particular, the sample complexity to recover K coordinate dictionaries with dimensions $m_k \times p_k$ up to estimation error ε_k is shown to be $\max_{k\in[K]} \mathcal{O}(m_k p_k^3 \varepsilon_k^{-2})$.**

***Index Terms*—Dictionary identification, dictionary learning, Kronecker-structured dictionary, sample complexity, sparse representations, tensor data, Tucker decomposition.**

I. Introduction

RAPID advances in sensing and data acquisition technologies are increasingly resulting in individual data samples or signals structured by multiple *modes*. Examples include hyperspectral video (four modes; two spatial, one temporal, and one spectral), colored depth video (five modes; two spatial, one temporal, one spectral, and one depth), and four-dimensional tomography (four modes; three spatial and one temporal). Such data form multiway arrays and are called *tensor data* [2], [3].

Typical feature extraction approaches that handle tensor data tend to collapse or vectorize the tensor into a long one-dimensional vector and apply existing processing methods for one-dimensional data. Such approaches ignore the structure and inter-mode correlations in tensor data. More recently, several works instead assume a structure on the tensor of interest through tensor decompositions such as the CANDECOMP/PARAFAC (CP) decomposition [4], Tucker decomposition [5], and PARATUCK decomposition [3] to obtain meaningful representations of tensor data. Because these decompositions involve fewer parameters, or degrees of freedom, in the model, inference algorithms that exploit such decompositions often perform better than those that assume the tensors to be unstructured. Moreover, algorithms utilizing tensor decompositions tend to be more efficient in terms of storage and computational costs: the cost of storing the decomposition can be substantially lower, and numerical methods can exploit the structure by solving simpler subproblems.

In this work, we focus on the problem of finding sparse representations of tensors that admit a Tucker decomposition. More specifically, we analyze the *dictionary learning* (DL) problem for tensor data. The traditional DL problem for vector-valued data involves constructing an overcomplete basis (dictionary) such that each data sample can be represented by only a few columns (atoms) of that basis [6]. To account for the Tucker structure of tensor data, we require that the dictionary underlying the vectorized versions of tensor data samples be *Kronecker structured* (KS). That is, it is comprised of *coordinate dictionaries* that independently transform various modes of the tensor data. Such dictionaries have successfully been used for tensor data representation in applications such as hyperspectral imaging, video acquisition, distributed sensing, magnetic resonance imaging, and the tensor completion problem (multidimensional inpainting) [7], [8]. To provide some insights into the usefulness of KS dictionaries for tensor data, consider the hypothetical problem of finding sparse representations of $1024 \times 1024 \times 32$ hyperspectral images. Traditional DL methods require each image to be rearranged into a one-dimensional vector of length 2^{25} and then learn an unstructured dictionary that has a total of $(2^{25}p)$ unknown parameters, where $p \geq 2^{25}$. In contrast, KS DL only requires learning three coordinate dictionaries of dimensions $1024 \times p_1$, $1024 \times p_2$, and $32 \times p_3$, where $p_1, p_2 \geq 1024$, and $p_3 \geq 32$. This gives rise to a total of $[1024(p_1 + p_2) + 32p_3]$ unknown parameters in KS DL, which is significantly smaller than $2^{25}p$. While such "parameter counting" points to the usefulness of KS DL for tensor data, a fundamental question remains open in the literature: what are the theoretical limits on the learning of KS dictionaries underlying Kth-order tensor data? To answer this question, we examine the KS-DL objective function and find sufficient conditions on the number of samples (or sample complexity) for successful local identification of *coordinate dictionaries* underlying the KS dictionary. To the best of our knowledge, this is the first work presenting such identification results for the KS-DL problem.

Manuscript received December 8, 2017; revised April 10, 2018; accepted May 3, 2018. Date of publication May 18, 2018; date of current version September 27, 2018. This work was supported in part by the National Science Foundation under Awards CCF-1525276 and CCF-1453073 and in part by the Army Research Office under Award W911NF-17-1-0546. This paper was presented in part at the 7th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, Curaçao, Dutch Antilles, December 2017. The guest editor coordinating the review of this paper and approving it for publication was Prof. Vincent Y. F. Tan. *(Corresponding author: Zahra Shakeri)*

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

Digital Object Identifier 10.1109/JSTSP.2018.2838092

A. Our Contributions

We derive sufficient conditions on the true coordinate dictionaries, coefficient and noise distributions, regularization parameter, and the number of data samples such that the KS-DL objective function has a local minimum within a small neighborhood of the true coordinate dictionaries with high probability. Specifically, suppose the observations are generated from a true dictionary $\mathbf{D}^0 \in \mathbb{R}^{m \times p}$ consisting of the Kronecker product of K coordinate dictionaries, $\mathbf{D}_k^0 \in \mathbb{R}^{m_k \times p_k}, k \in \{1, \ldots, K\}$, where $m = \prod_{k=1}^K m_k$ and $p = \prod_{k=1}^K p_k$. Our results imply that $N = \max_{k \in [K]} \Omega(m_k p_k^3 \varepsilon_k^{-2})$ samples are sufficient (with high probability) to recover the underlying coordinate dictionaries $\mathbf{D}_k^0$ up to the given estimation errors $\varepsilon_k, k \in \{1, \ldots, K\}$.

B. Relationship to Prior Work

Among existing works on structured DL that have focused exclusively on the Tucker model for tensor data, several have only empirically established the superiority of KS DL in various settings for 2nd and 3rd-order tensor data [8]–[13].

In the case of unstructured dictionaries, several works do provide analytical results for the dictionary identifiability problem [14]–[21]. These results, which differ from each other in terms of the distance metric used, cannot be trivially extended for the KS-DL problem. In this work, we focus on the Frobenius norm as the distance metric. Gribonval *et al.* [20] and Jung *et al.* [21] also consider this metric, with the latter work providing minimax lower bounds for dictionary reconstruction error. In particular, Jung *et al.* [21] show that the number of samples needed for reliable reconstruction (up to a prescribed mean squared error ε) of an $m \times p$ dictionary within its local neighborhood must be *at least* on the order of $N = \Omega(mp^2\varepsilon^{-2})$. Gribonval *et al.* [20] derive a competing upper bound for the sample complexity of the DL problem and show that $N = \Omega(mp^3\varepsilon^{-2})$ samples are *sufficient* to guarantee (with high probability) the existence of a local minimum of the DL cost function within the ε neighborhood of the true dictionary. In our previous works, we have obtained lower bounds on the minimax risk of KS DL for 2nd-order [22] and Kth-order tensors [23], [24], and have shown that the number of samples necessary for reconstruction of the true KS dictionary within its local neighborhood up to a given estimation error scales with the sum of the product of the dimensions of the coordinate dictionaries, i.e., $N = \Omega(p \sum_{k=1}^K m_k p_k \varepsilon^{-2})$. Compared to this sample complexity lower bound, our upper bound is larger by a factor $\max_k p_k^2$.

In terms of the analytical approach, although we follow the same general proof strategy as the vectorized case of Gribonval *et al.* [20], our extension poses several technical challenges. These include: (i) expanding the asymptotic objective function into a summation in which individual terms depend on coordinate dictionary recovery errors, (ii) translating identification conditions on the KS dictionary to conditions on its coordinate dictionaries, and (iii) connecting the asymptotic objective function to the empirical objective function using concentration of measure arguments; this uses the *coordinate-wise Lipschitz continuity* property of the KS-DL objective function with respect to the coordinate dictionaries. To address these challenges, we require additional assumptions on the generative model. These include: (i) the true dictionary and the recovered dictionary belong to the class of KS dictionaries, and (ii) dictionary coefficient tensors follow the *separable sparsity* model that requires nonzero coefficients to be grouped in blocks [24], [25].

C. Notational Convention and Preliminaries

Underlined bold upper-case, bold upper-case and lower-case letters are used to denote tensors, matrices and vectors, respectively, while non-bold lower-case letters denote scalars. For a tensor $\underline{\mathbf{X}}$, its $(i_1, \ldots, i_K)$-th element is denoted as $\underline{x}_{i_1 \ldots i_K}$. The i-th element of vector $\mathbf{v}$ is denoted by v_i and the ij-th element of matrix $\mathbf{X}$ is denoted as x_{ij}. The k-th column of $\mathbf{X}$ is denoted by $\mathbf{x}_k$ and $\mathbf{X}_{\mathcal{I}}$ denotes the matrix consisting of the columns of $\mathbf{X}$ with indices $\mathcal{I}$. We use $|\mathcal{I}|$ for the cardinality of the set $\mathcal{I}$. Sometimes we use matrices indexed by numbers, such as $\mathbf{X}_1$, in which case a second index (e.g., $\mathbf{x}_{1,k}$) is used to denote its columns. 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 use $\mathrm{diag}\,(\mathbf{X})$ to denote the vector comprised of the diagonal elements of $\mathbf{X}$ and $\mathrm{Diag}\,(\mathbf{v})$ to denote the diagonal matrix, whose diagonal elements are comprised of elements of $\mathbf{v}$. The elements of the sign vector of $\mathbf{v}$, denoted as $\mathrm{sign}(\mathbf{v})$, are equal to $\mathrm{sign}(v_i) = v_i/|v_i|$, for $v_i \neq 0$, and $\mathrm{sign}(v_i) = 0$ for $v_i = 0$, where i denotes the index of any element of v. We also use $\sin(\mathbf{v})$ to denote the vector with elements $\sin(v_i)$ (used similarly for other trigonometric functions). Norms are given by subscripts, so $\|\mathbf{v}\|_0$, $\|\mathbf{v}\|_1$, and $\|\mathbf{v}\|_2$ are the ℓ_0, ℓ_1, and ℓ_2 norms of $\mathbf{v}$, while $\|\mathbf{X}\|_2$ and $\|\mathbf{X}\|_F$ are the spectral and Frobenius norms of $\mathbf{X}$, respectively. We use $[K]$ to denote $\{1, 2, \ldots, K\}$ and $\mathbf{X}_{1:K}$ to denote $\{\mathbf{X}_k\}_{k=1}^K$.

We write $\mathbf{X} \otimes \mathbf{Y}$ for the *Kronecker product* of two matrices $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\mathbf{Y} \in \mathbb{R}^{p \times q}$, where the result is an $mp \times nq$ matrix and we have $\|\mathbf{X} \otimes \mathbf{Y}\|_F = \|\mathbf{X}\|_F \|\mathbf{Y}\|_F$ [26]. We also use $\bigotimes_{k \in K} \mathbf{X}_k \triangleq \mathbf{X}_1 \otimes \cdots \otimes \mathbf{X}_K$. We define $\mathbf{H}_{\mathbf{X}} \triangleq (\mathbf{X}^\top \mathbf{X})^{-1}$, $\mathbf{X}^+ \triangleq \mathbf{H}_{\mathbf{X}} \mathbf{X}^\top$, and $\mathbf{P}_{\mathbf{X}} \triangleq \mathbf{X}\mathbf{X}^+$ for full rank matrix $\mathbf{X}$. In the body, we sometimes also use $\Delta f(\mathbf{X}; \mathbf{Y}) \triangleq f(\mathbf{X}) - f(\mathbf{Y})$.

For matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ of appropriate dimensions, we define their distance to be $d(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|_F$. For $\mathbf{X}^0$ belonging to some set $\mathcal{X}$, we define

$$\begin{aligned} \mathcal{S}_\varepsilon(\mathbf{X}^0) &\triangleq \left\{\mathbf{X} \in \mathcal{X} : \|\mathbf{X} - \mathbf{X}^0\|_F = \varepsilon\right\}, \\ \mathcal{B}_\varepsilon(\mathbf{X}^0) &\triangleq \left\{\mathbf{X} \in \mathcal{X} : \|\mathbf{X} - \mathbf{X}^0\|_F < \varepsilon\right\}, \\ \bar{\mathcal{B}}_\varepsilon(\mathbf{X}^0) &\triangleq \left\{\mathbf{X} \in \mathcal{X} : \|\mathbf{X} - \mathbf{X}^0\|_F \leq \varepsilon\right\}. \end{aligned} \tag{1}$$

Note that while $\mathcal{S}_\varepsilon(\mathbf{X}^0)$ represents the surface of a sphere, we use the term "sphere" for simplicity. We use the standard "big-$\mathcal{O}$" (Knuth) notation for asymptotic scaling.

1) Tensor Operations and Tucker Decomposition for Tensors: A tensor is a multidimensional array where the order of the tensor is defined as the number of dimensions in the array.

Tensor Unfolding: 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 k th-order. The k-rank of a tensor $\underline{\mathbf{X}}$ is defined by $\operatorname{rank}(\mathbf{X}_{(k)})$; trivially, $\operatorname{rank}(\mathbf{X}_{(k)}) \leq p_k$.

Tensor Multiplication: 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 $(\underline{\mathbf{X}} \times_k \mathbf{A})_{i_1 \ldots i_{k-1} j i_{k+1} \ldots i_K} = \sum_{i_k=1}^{p_k} \underline{x}_{i_1 \ldots i_{k-1} i_k i_{k+1} \ldots i_K} a_{j i_k}$. 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 [3]:

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_k \mathbf{A} \Leftrightarrow \mathbf{Y}_{(k)} = \mathbf{A}\mathbf{X}_{(k)}. \tag{2}$$

Tucker Decomposition: The Tucker decomposition decomposes a tensor into a *core tensor* multiplied by a matrix along each mode [3], [5]. We take advantage of the Tucker model since we can relate the Tucker decomposition to the Kronecker representation of tensors [25]. For a tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$ of order K, if $\operatorname{rank}(\mathbf{Y}_{(k)}) \leq p_k$ holds for all $k \in [K]$ then, according to the Tucker model, $\underline{\mathbf{Y}}$ can be decomposed into:

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times_3 \cdots \times_K \mathbf{D}_K, \tag{3}$$

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. The following is implied by (3) [3]:

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

Since the Kronecker product satisfies $\operatorname{vec}(\mathbf{B}\mathbf{X}\mathbf{A}^\top) = (\mathbf{A} \otimes \mathbf{B}) \operatorname{vec}(\mathbf{X})$, (3) is equivalent to

$$\operatorname{vec}(\underline{\mathbf{Y}}) = \left(\mathbf{D}_K \otimes \mathbf{D}_{K-1} \otimes \cdots \otimes \mathbf{D}_1\right) \operatorname{vec}(\underline{\mathbf{X}}), \tag{4}$$

where $\operatorname{vec}(\underline{\mathbf{Y}}) \triangleq \operatorname{vec}(\mathbf{Y}_{(1)})$ and $\operatorname{vec}(\underline{\mathbf{X}}) \triangleq \operatorname{vec}(\mathbf{X}_{(1)})$.

2) Definitions for Matrices: We use the following definitions for a matrix $\mathbf{D}$ with unit-norm columns: $\delta_s(\mathbf{D})$ denotes the *restricted isometry property* (RIP) constant of order s for $\mathbf{D}$ [27]. We define the *worst-case coherence* of $\mathbf{D}$ as $\mu_1(\mathbf{D}) = \max_{\substack{i,j \\ i \neq j}} \left|\mathbf{d}_i^\top \mathbf{d}_j\right|$. We also define the *order-s cumulative coherence* of $\mathbf{D}$ as

$$\mu_s(\mathbf{D}) \triangleq \max_{|\mathcal{J}| \leq s} \max_{j \notin \mathcal{J}} \|\mathbf{D}_{\mathcal{J}}^\top \mathbf{d}_j\|_1. \tag{5}$$

Note that for $s = 1$, the cumulative coherence is equivalent to the worst-case coherence and $\mu_s(\mathbf{D}) \leq s\mu_1(\mathbf{D})$ [20]. For $\mathbf{D} = \bigotimes_{k \in [K]} \mathbf{D}_k$, where $\mathbf{D}_k$'s have unit-norm columns, $\mu_1(\mathbf{D}) = \max_{k \in [K]} \mu_1(\mathbf{D}_k)$ [28, Corollary 3.6] and it can be shown that[1]:

$$\mu_s(\mathbf{D}) \leq \max_{k \in [K]} \mu_{s_k}(\mathbf{D}_k) \left(\prod_{\substack{i \in [K], \\ i \neq k}} (1 + \mu_{s_i - 1}(\mathbf{D}_i)) \right). \tag{6}$$

The rest of the paper is organized as follows. We formulate the KS-DL problem in Section II. In Section III, we provide analysis for asymptotic recovery of coordinate dictionaries composing the KS dictionary and in Section IV, we present sample complexity results for identification of coordinate dictionaries that are based on the results of Section III. Finally, we conclude the paper in Section V. In order to keep the main exposition simple, proofs of the lemmas and propositions are relegated to appendices.

[1]The proof of (6) is provided in Appendix C.

II. System Model

We assume the observations are Kth-order tensors $\underline{\mathbf{Y}} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$. Given generating *coordinate dictionaries* $\mathbf{D}_k^0 \in \mathbb{R}^{m_k \times p_k}$, *coefficient tensor* $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_K}$, and *noise tensor* $\underline{\mathbf{W}}$, we can write $\mathbf{y} \triangleq \operatorname{vec}(\underline{\mathbf{Y}})$ using (4) as[2]

$$\mathbf{y} = \left(\bigotimes_{k \in [K]} \mathbf{D}_k^0 \right) \mathbf{x} + \mathbf{w}, \quad \|\mathbf{x}\|_0 \leq s, \tag{7}$$

where $\mathbf{x} = \operatorname{vec}(\underline{\mathbf{X}}) \in \mathbb{R}^p$ denotes the sparse generating coefficient vector, $\mathbf{D}^0 = \bigotimes \mathbf{D}_k^0 \in \mathbb{R}^{m \times p}$ denotes the underlying KS dictionary, and $\mathbf{w} = \operatorname{vec}(\underline{\mathbf{W}}) \in \mathbb{R}^m$ denotes the underlying noise vector. Here, $\mathbf{D}_k^0 \in \mathcal{D}_k = \{\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}, \|\mathbf{d}_{k,j}\|_2 = 1, \forall j \in [p_k]\}$ for $k \in [K]$, $p = \prod_{k \in [K]} p_k$ and $m = \prod_{k \in [K]} m_k$.[3] We use $\bigotimes$ for $\bigotimes_{k \in [K]}$ in the following for simplicity of notation. We assume we are given N noisy tensor observations, which are then stacked in a matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$. To state the problem formally, we first make the following assumptions on distributions of $\mathbf{x}$ and $\mathbf{w}$ for each tensor observation.

Coefficient distribution: We assume the coefficient tensor $\underline{\mathbf{X}}$ follows the random *"separable sparsity"* model. That is, $\mathbf{x} = \operatorname{vec}(\underline{\mathbf{X}})$ is sparse and the support of nonzero entries of $\mathbf{x}$ is structured and random. Specifically, we sample s_k elements uniformly at random from $[p_k]$, $k \in [K]$. Then, the random support of $\mathbf{x}$ is $\{\mathcal{J} \subseteq [p], |\mathcal{J}| = s\}$ and is associated with

$$\{\mathcal{J}_1 \times \mathcal{J}_2 \times \cdots \times \mathcal{J}_K : \mathcal{J}_k \subseteq [p_k], |\mathcal{J}_k| = s_k, k \in [K]\}$$

via lexicographic indexing, where $s = \prod_{k \in [K]} s_k$, and the support of $\mathbf{x}_{1:N}$'s are assumed to be independent and identically distributed (i.i.d.). This model requires nonzero entries of the coefficient tensors to be grouped in blocks and the sparsity level associated with each coordinate dictionary to be small [25].[4]

We now make the same assumptions for the distribution of $\mathbf{x}$ as assumptions A and B in Gribonval *et al.* [20]. These include: (i) $\mathbb{E}\left\{\mathbf{x}_{\mathcal{J}}\mathbf{x}_{\mathcal{J}}^\top | \mathcal{J}\right\} = \mathbb{E}\left\{x^2\right\}\mathbf{I}_s$, (ii) $\mathbb{E}\left\{\mathbf{x}_{\mathcal{J}}\boldsymbol{\sigma}_{\mathcal{J}}^\top | \mathcal{J}\right\} = \mathbb{E}\left\{|x|\right\}\mathbf{I}_s$, where $\boldsymbol{\sigma} = \operatorname{sign}(\mathbf{x})$, (iii) $\mathbb{E}\left\{\boldsymbol{\sigma}_{\mathcal{J}}\boldsymbol{\sigma}_{\mathcal{J}}^\top | \mathcal{J}\right\} = \mathbf{I}_s$, (iv) magnitude of $\mathbf{x}$ is bounded, i.e., $\|\mathbf{x}\|_2 \leq M_x$ almost surely, and (v) nonzero entries of $\mathbf{x}$ have a minimum magnitude, i.e., $\min_{j \in \mathcal{J}} |x_j| \geq x_{\min}$ almost surely. Finally, we define $\kappa_x \triangleq \mathbb{E}\left\{|x|\right\} / \sqrt{\mathbb{E}\left\{x^2\right\}}$ as a measure of the flatness of $\mathbf{x}$ ($\kappa_x \leq 1$, with $\kappa_x = 1$ when all nonzero coefficients are equal [20]).

Noise distribution: We make following assumptions on the distribution of noise, which is assumed i.i.d. across data samples: (i) $\mathbb{E}\left\{\mathbf{w}\mathbf{w}^\top\right\} = \mathbb{E}\left\{w^2\right\}\mathbf{I}_m$, (ii) $\mathbb{E}\left\{\mathbf{w}\mathbf{x}^\top | \mathcal{J}\right\} =$

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

[3]Note that the $\mathcal{D}_k$'s are compact sets on their respective oblique manifolds of matrices with unit-norm columns [20].

[4]In contrast, for coefficients following the random non-separable sparsity model, the support of the nonzero entries of the coefficient vector are assumed uniformly distributed over $\{\mathcal{J} \subseteq [p] : |\mathcal{J}| = s\}$.

$\mathbb{E}\left\{\mathbf{w}\boldsymbol{\sigma}^\top|\mathcal{J}\right\}=\mathbf{0}$, and (iii) magnitude of $\mathbf{w}$ is bounded, i.e., $\|\mathbf{w}\|_2 \leq M_w$ almost surely.

Our goal in this paper is to recover the underlying coordinate dictionaries, $\mathbf{D}_k^0$, from N noisy realizations of tensor data. To solve this problem, we take the empirical risk minimization approach and define

$$f_{\mathbf{y}}\left(\mathbf{D}_{1:K}\right) \triangleq \inf_{\mathbf{x}'\in\mathbb{R}^p}\left\{\frac{1}{2}\left\|\mathbf{y}-\left(\bigotimes \mathbf{D}_k\right)\mathbf{x}'\right\|_2^2+\lambda\|\mathbf{x}'\|_1\right\}, \text{and}$$

$$F_{\mathbf{Y}}\left(\mathbf{D}_{1:K}\right) \triangleq \frac{1}{N}\sum_{n=1}^{N} f_{\mathbf{y}_n}\left(\mathbf{D}_{1:K}\right), \tag{8}$$

where λ is a regularization parameter. In theory, we can recover the coordinate dictionaries by solving the following regularized optimization program:

$$\min_{\substack{\mathbf{D}_k\in\mathcal{D}_k\\ k\in[K]}} F_{\mathbf{Y}}\left(\mathbf{D}_{1:K}\right). \tag{9}$$

More specifically, given desired errors $\{\varepsilon_k\}_{k=1}^K$, we want a local minimum of (9) to be attained by coordinate dictionaries $\widehat{\mathbf{D}}_k \in \mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0), k\in[K]$. That is, there exists a set $\{\widehat{\mathbf{D}}_k\}_{k\in[K]} \subset \left\{\mathbf{D}_k \in \mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0)\right\}_{k\in[K]}$ such that $F_{\mathbf{Y}}(\widehat{\mathbf{D}}_{1:K}) \leq F_{\mathbf{Y}}(\mathbf{D}_{1:K})$.[5] To address this problem, we first minimize the statistical risk:

$$\min_{\substack{\mathbf{D}_k\in\mathcal{D}_k\\ k\in[K]}} f_{\mathbb{P}}\left(\mathbf{D}_{1:K}\right) \triangleq \min_{\substack{\mathbf{D}_k\in\mathcal{D}_k\\ k\in[K]}} \mathbb{E}_{\mathbf{y}}\left\{f_{\mathbf{y}}\left(\mathbf{D}_{1:K}\right)\right\}. \tag{10}$$

Then, we connect $F_{\mathbf{Y}}\left(\mathbf{D}_{1:K}\right)$ to $f_{\mathbb{P}}\left(\mathbf{D}_{1:K}\right)$ using concentration of measure arguments and obtain the number of samples sufficient for local recovery of the coordinate dictionaries. Such a result ensures that any KS-DL algorithm that is guaranteed to converge to a local minimum, and which is initialized close enough to the true KS dictionary, will converge to a solution close to the generating coordinate dictionaries (as opposed to the generating KS dictionary, which is guaranteed by analysis of the vector-valued setup [20]).

III. Asympototic Identifiability Results

In this section, we provide an identifiability result for the KS-DL objective function in (10). The implications of this theorem are discussed in Section V.

Theorem 1: Suppose the observations are generated according to (7) and the dictionary coefficients follow the separable sparsity model of Section II. Further, assume the following conditions are satisfied:

$$s_k \leq \frac{p_k}{8\left(\|\mathbf{D}_k^0\|_2+1\right)^2},$$

$$\max_{k\in[K]}\left\{\mu_{s_k}(\mathbf{D}_k^0)\right\} \leq \frac{1}{4}, \quad \mu_s(\mathbf{D}^0) < \frac{1}{2}, \tag{11}$$

and

$$\frac{\mathbb{E}\left\{x^2\right\}}{M_x\mathbb{E}\left\{|x|\right\}} > \frac{24\sqrt{3}(4.5^{K/2})K}{(1-2\mu_s(\mathbf{D}^0))} \max_{k\in[K]}\left\{\frac{s_k}{p_k}\left\|\mathbf{D}_k^{0\top}\mathbf{D}_k^0-\mathbf{I}\right\|_F\left(\|\mathbf{D}_k^0\|_2+1\right)\right\}. \tag{12}$$

Define

$$C_{k,\min} \triangleq 8(3^{\frac{K+1}{2}})\kappa_x^2\left(\frac{s_k}{p_k}\right)\left\|\mathbf{D}_k^{0\top}\mathbf{D}_k^0-\mathbf{I}\right\|_F\left(\|\mathbf{D}_k^0\|_2+1\right),$$

$$C_{\max} \triangleq \frac{1}{3K(1.5)^{K/2}}\frac{\mathbb{E}\left\{|x|\right\}}{M_x}(1-2\mu_s(\mathbf{D}^0)). \tag{13}$$

Then, the map $\mathbf{D}_{1:K} \mapsto f_{\mathbb{P}}\left(\mathbf{D}_{1:K}\right)$ admits a local minimum $\widehat{\mathbf{D}} = \bigotimes_{k\in[K]}\widehat{\mathbf{D}}_k$ such that $\widehat{\mathbf{D}}_k \in \mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0)$, $k\in[K]$, for any $\varepsilon_k > 0$ as long as

$$\lambda \leq \frac{x_{\min}}{8\times 3^{(K-1)/2}}, \tag{14}$$

$$\frac{\lambda C_{k,\min}}{\mathbb{E}\left\{|x|\right\}} < \varepsilon_k < \frac{\lambda C_{\max}}{\mathbb{E}\left\{|x|\right\}}, \ k\in[K], \tag{15}$$

and

$$\frac{M_w}{M_x} < 3(1.5)^{K/2}\left(\frac{\lambda K C_{\max}}{\mathbb{E}\left\{|x|\right\}} - \sum_{k\in[K]}\varepsilon_k\right). \tag{16}$$

A. Discussion

Theorem 1 captures how the existence of a local minimum for the statistical risk minimization problem depends on various properties of the coordinate dictionaries and demonstrates that there exists a local minimum of $f_{\mathbb{P}}\left(\mathbf{D}_{1:K}\right)$ that is in local neighborhoods of the coordinate dictionaries. This ensures asymptotic recovery of coordinate dictionaries within some local neighborhood of the true coordinate dictionaries, as opposed to KS dictionary recovery for vectorized observations [20, Th. 1].

We now explicitly compare conditions in Theorem 1 with the corresponding ones for vectorized observations [20, Th. 1]. Given that the coefficients are drawn from the separable sparsity model, the sparsity constraints for the coordinate dictionaries in (11) translate into

$$\frac{s}{p} = \prod_{k\in[K]}\frac{s_k}{p_k} \leq \frac{1}{8^K\prod_k\left(\|\mathbf{D}_k^0\|_2+1\right)^2}. \tag{17}$$

Therefore, we have $\frac{s}{p} = \mathcal{O}(\frac{1}{\prod_k\|\mathbf{D}_k^0\|_2^2}) = \mathcal{O}(\frac{1}{\|\mathbf{D}^0\|_2^2})$. Using the fact that $\|\mathbf{D}^0\|_2 \geq \|\mathbf{D}^0\|_F/\sqrt{m} = \sqrt{p}/\sqrt{m}$, this translates into sparsity order $s = \mathcal{O}(m)$. Next, the left hand side of the condition in (12) is less than 1. Moreover, from properties of the Frobenius norm, it is easy to show that $\|\mathbf{D}_k^{0\top}\mathbf{D}_k^0-\mathbf{I}\|_F \geq \sqrt{p_k(p_k-m_k)/m_k}$. The fact that $\left\|\mathbf{D}_k^0\right\|_2 \geq \sqrt{p_k}/\sqrt{m_k}$ and the assumption $\mu_{s_k}(\mathbf{D}_k^0) \leq 1/4$ imply that the right hand side of (12) is lower bounded by $\Omega(\max_k s_k\sqrt{(p_k-m_k)/m_k^2})$. Therefore, Theorem 1 applies to coordinate dictionaries with dimensions $p_k \leq m_k^2$ and subsequently, KS dictionaries with $p \leq m^2$.

[5] We focus on the local recovery of coordinate dictionaries (i.e., $\widehat{\mathbf{D}}_k \in \mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0)$) due to ambiguities in the general DL problem. This ambiguity is a result of the fact that dictionaries are invariant to permutation and sign flips of dictionary columns, resulting in equivalent classes of dictionaries. Some works in the literature on conventional DL overcome this issue by defining distance metrics that capture the distance between these equivalent classes [15]–[17].

Both the sparsity order and dictionary dimensions are in line with the scaling results for vectorized data [20].

B. Proof Outline

For given radii $0 < \varepsilon_k \leq 2\sqrt{p_k}, k \in [K]$, the spheres $\mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)$ are non-empty. This follows from the construction of dictionary classes, $\mathcal{D}_k$'s. Moreover, the mapping $\mathbf{D}_{1:K} \mapsto f_{\mathbb{P}}(\mathbf{D}_{1:K})$ is continuous with respect to the Frobenius norm $\|\mathbf{D}_k - \mathbf{D}_k'\|_F$ on all $\mathbf{D}_k, \mathbf{D}_k' \in \mathbb{R}^{m_k \times p_k}, k \in [K]$ [29]. Hence, it is also continuous on compact constraint sets $\mathcal{D}_k$'s. We derive conditions on the coefficients, underlying coordinate dictionaries, M_w, regularization parameter, and ε_k's such that

$$\Delta f_{\mathbb{P}}(\varepsilon_{1:K}) \triangleq \inf_{\mathbf{D}_k \in \mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)} \Delta f_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0) > 0. \quad (18)$$

This along with the compactness of closed balls $\bar{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0)$ and the continuity of the mapping $\mathbf{D}_{1:K} \mapsto f_{\mathbb{P}}(\mathbf{D}_{1:K})$ imply the existence of a local minimum of $f_{\mathbb{P}}(\mathbf{D}_{1:K})$ achieved by $\widehat{\mathbf{D}}_{1:K}$ in open balls, $\mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0)$'s, $k \in [K]$.

To find conditions that ensure $\Delta f_{\mathbb{P}}(\varepsilon_{1:K}) > 0$, we take the following steps: given coefficients that follow the separable sparsity model, we can decompose any $\mathbf{D}_{\mathcal{J}}, |\mathcal{J}| = s$, as

$$\mathbf{D}_{\mathcal{J}} = \bigotimes \mathbf{D}_{k,\mathcal{J}_k}, \quad (19)$$

where $|\mathcal{J}_k| = s_k$ for $k \in [K]$.[6] Given a generating $\boldsymbol{\sigma} = \operatorname{sign}(\mathbf{x})$, we obtain $\widehat{\mathbf{x}}$ by solving $f_{\mathbf{y}}(\mathbf{D}_{1:K})$ with respect to $\mathbf{x}'$, conditioned on the fact that $\operatorname{sign}(\widehat{\mathbf{x}}) = \widehat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}$. This eliminates the dependency of $f_{\mathbf{y}}(\mathbf{D}_{1:K})$ on $\inf_{\mathbf{x}'}$ by finding a closed-form expression for $f_{\mathbf{y}}(\mathbf{D}_{1:K})$ given $\widehat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}$, which we denote as $\phi_{\mathbf{y}}(\mathbf{D}_{1:K}|\boldsymbol{\sigma})$. Defining

$$\phi_{\mathbb{P}}(\mathbf{D}_{1:K}|\boldsymbol{\sigma}) \triangleq \mathbb{E}\{\phi_{\mathbf{y}}(\mathbf{D}_{1:K}|\boldsymbol{\sigma})\}, \quad (20)$$

we expand $\Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma})$ using (19) and separate the terms that depend on each radius $\varepsilon_k = \|\mathbf{D}_k - \mathbf{D}_k^0\|_F$ to obtain conditions for sparsity levels $s_k, k \in [K]$, and coordinate dictionaries such that $\Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}) > 0$. Finally, we derive conditions on M_w, coordinate dictionary coherences and ε_k's that ensure $\widehat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}$ and $\Delta f_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0) = \Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma})$.

Remark 1: The key assumption in the proof of Theorem 1 is expanding $\mathbf{D}_{\mathcal{J}}$ according to (19). This is a consequence of the separable sparsity model for dictionary coefficients. For a detailed discussion on the differences between the separable sparsity model and the random sparsity model for tensors, we refer the readers to our earlier work [22].

Remark 2: Although some of the forthcoming lemmas needed of Theorem 1 impose conditions on $\mathbf{D}_k$'s as well as true coordinate dictionaries $\mathbf{D}_k^0$'s, we later translate these conditions exclusively in terms of $\mathbf{D}_k^0$'s and ε_k's.

The proof of Theorem 1 relies on the following propositions and lemmas. The proofs of these are provided in Appendix A.

[6]The separable sparsity distribution model implies sampling without replacement from columns of $\mathbf{D}_k$.

Proposition 1: Suppose the following inequalities hold for $k \in [K]$:

$$s_k \leq \frac{p_k}{8(\|\mathbf{D}_k^0\|_2 + 1)^2} \quad \text{and} \quad \max_{k \in [K]} \{\delta_{s_k}(\mathbf{D}_k^0)\} \leq \frac{1}{4}. \quad (21)$$

Then, for

$$\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\{|x|\}} \leq \frac{1}{8 \times 3^{(K-1)/2}}, \quad (22)$$

any collection of $\{\varepsilon_k : \varepsilon_k \leq 0.15, k \in [K]\}$, and for all $\mathbf{D}_k \in \mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)$, we have :

$$\Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}) \geq \frac{s\mathbb{E}\{x^2\}}{8} \sum_{k \in [K]} \frac{\varepsilon_k}{p_k} (\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda})), \quad (23)$$

where

$$\varepsilon_{k,\min}(\bar{\lambda}) \triangleq \frac{3^{(K-1)/2}}{2} \left(1.5^{\frac{K-1}{2}} + 2^{(K+1)}\bar{\lambda}\right) \bar{\lambda} C_{k,\min}.$$

In addition, if

$$\bar{\lambda} \leq \frac{0.15}{\max_{k \in [K]} C_{k,\min}}, \quad (24)$$

then $\varepsilon_{k,\min}(\bar{\lambda}) < 0.15$. Thus, $\Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}) > 0$ for all $\varepsilon_k \in (\varepsilon_{k,\min}(\bar{\lambda}), 0.15], k \in [K]$.

The proof of Proposition 1 relies on the following lemmas as well as supporting lemmas from the analysis of vectorized data [20, Lemmas 4,6,7,15,16].

Lemma 1: Let $\mathbf{D} = \bigotimes \mathbf{D}_k$ where $\delta_s(\mathbf{D}_k) < 1$ for $k \in [K]$, and $\mathcal{J}$ be a support set generated by the separable sparsity model. Then any $\mathbf{D}_{\mathcal{J}}, |\mathcal{J}| = s$, can be decomposed as $\mathbf{D}_{\mathcal{J}} = \bigotimes \mathbf{D}_{k,\mathcal{J}_k}$, where $|\mathcal{J}_k| = s_k$ and $\operatorname{rank}(\mathbf{D}_{k,\mathcal{J}_k}) = s_k$, for $k \in [K]$. Also, the following relations hold for this model:[7]

$$\mathbf{P}_{\mathbf{D}_{\mathcal{J}}} = \bigotimes \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}, \mathbf{D}_{\mathcal{J}}^+ = \bigotimes \mathbf{D}_{k,\mathcal{J}_k}^+, \mathbf{H}_{\mathbf{D}_{\mathcal{J}}} = \bigotimes \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}, \quad (25)$$

where $\mathbf{P}$ and $\mathbf{H}$ are defined in Section I-C.

Lemma 2: Given $\mathbf{D}_{1:K}$ and $\mathbf{D}_{1:K}^0$, the difference

$$\bigotimes \mathbf{D}_k - \bigotimes \mathbf{D}_k^0 = \sum_{k \in [K]} \widetilde{\mathbf{D}}_{k,1} \otimes \cdots \otimes (\mathbf{D}_k - \mathbf{D}_k^0) \otimes \cdots \otimes \widetilde{\mathbf{D}}_{k,K}, \quad (26)$$

where without loss of generality, each $\widetilde{\mathbf{D}}_{k,i}$ is equal to either $\mathbf{D}_i^0$ or $\mathbf{D}_i$, for $k \in [K]$.

We drop the k index from $\widetilde{\mathbf{D}}_{k,i}$ for ease of notation throughout the rest of the paper.

Lemma 3: Let $\boldsymbol{\sigma} \in \{-1, 0, 1\}^p$ be an arbitrary sign vector and $\mathcal{J} = \mathcal{J}(\boldsymbol{\sigma})$ be its support. Define[8]

$$\phi_{\mathbf{y}}(\mathbf{D}_{1:K}|\boldsymbol{\sigma}) \triangleq \inf_{\substack{\mathbf{x} \in \mathbb{R}^p \\ \operatorname{supp}(\mathbf{x}) \subset \mathcal{J}}} \frac{1}{2} \left\| \mathbf{y} - \left(\bigotimes \mathbf{D}_k\right) \mathbf{x} \right\|_2^2 + \lambda \boldsymbol{\sigma}^\top \mathbf{x}. \quad (27)$$

[7]The equations follow from basic properties of the Kronecker product [26].

[8]The quantity $\phi_{\mathbf{y}}(\mathbf{D}_{1:K}|\boldsymbol{\sigma})$ is not equal to $\phi_{\mathbf{y}}(\mathbf{D}_{1:K})$ conditioned on $\boldsymbol{\sigma}$ and the expression is only used for notation.

If $\mathbf{D}_{k,\mathcal{J}_k}^\top \mathbf{D}_{k,\mathcal{J}_k}$ is invertible for $k \in [K]$, then $\widehat{\mathbf{x}}$ minimizes $\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right)$, where

$$\widehat{\mathbf{x}}_{\mathcal{J}} = \left(\bigotimes \mathbf{D}_{k,\mathcal{J}_k}^{+}\right)\mathbf{y} - \lambda\left(\bigotimes \left(\mathbf{D}_{k,\mathcal{J}_k}^\top \mathbf{D}_{k,\mathcal{J}_k}\right)^{-1}\right)\boldsymbol{\sigma}_{\mathcal{J}}, \quad (28)$$

and $\widehat{\mathbf{x}}_{\mathcal{J}^c} = \mathbf{0}$. Thus, $\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right)$ can be expressed in closed form as:

$$\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\mathbf{y}^\top\left(\bigotimes \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\mathbf{y} + \lambda\boldsymbol{\sigma}_{\mathcal{J}}^\top\left(\bigotimes \mathbf{D}_{k,\mathcal{J}_k}^{+}\right)\mathbf{y} - \frac{\lambda^2}{2}\boldsymbol{\sigma}_{\mathcal{J}}^\top\left(\bigotimes \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\boldsymbol{\sigma}_{\mathcal{J}}. \quad (29)$$

Lemma 4: Assume $\max\left\{\delta_{s_k}(\mathbf{D}_k^0), \delta_{s_k}(\mathbf{D}_k)\right\} < 1$ for $k \in [K]$ and let $\widetilde{\mathbf{D}}_k$ be equal to either $\mathbf{D}_k^0$ or $\mathbf{D}_k$. For

$$\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) \triangleq \phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right) - \phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right), \quad (30)$$

we have

$$\begin{aligned}
&\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\\
&= \frac{\mathbb{E}\{x^2\}}{2}\sum_{k\in[K]}\mathbb{E}_{\mathcal{J}_1}\left\{\mathrm{Tr}\left[\mathbf{D}_1^{0\top}\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\mathbf{D}_1^0\right]\right\}\cdots\\
&\qquad \mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{D}_k^{0\top}(\mathbf{I}_{m_k} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}})\mathbf{D}_k^0\right]\right\}\\
&\qquad \cdots\mathbb{E}_{\mathcal{J}_K}\left\{\mathrm{Tr}\left[\mathbf{D}_K^{0\top}\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\mathbf{D}_K^0\right]\right\}\\
&- \lambda\mathbb{E}\{|x|\}\sum_{k\in[K]}\mathbb{E}_{\mathcal{J}_1}\left\{\mathrm{Tr}\left[\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}^{+}\mathbf{D}_1^0\right]\right\}\cdots\\
&\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{I}_{s_k} - \mathbf{D}_{k,\mathcal{J}_k}^{+}\mathbf{D}_k^0\right]\right\}\cdots\mathbb{E}_{\mathcal{J}_K}\left\{\mathrm{Tr}\left[\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}^{+}\mathbf{D}_K^0\right]\right\}\\
&+ \frac{\lambda^2}{2}\sum_{k\in[K]}\mathbb{E}_{\mathcal{J}_1}\left\{\mathrm{Tr}\left[\mathbf{H}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\right]\right\}\cdots\\
&\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}^0} - \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right]\right\}\cdots\mathbb{E}_{\mathcal{J}_K}\left\{\mathrm{Tr}\left[\mathbf{H}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\right]\right\}.
\end{aligned} \quad (31)$$

Lemma 5: For any $\mathbf{D}_k \in \mathcal{D}_k$ satisfying RIP of order s_k, given $\mathcal{J}_k \subset [p_k]$ and $|\mathcal{J}_k| = s_k$, the following relations hold:

$$\|\mathbf{D}_{k,\mathcal{J}_k}\|_2 = \left\|\mathbf{D}_{k,\mathcal{J}_k}{}^\top\right\|_2 \le \sqrt{1+\delta_{s_k}(\mathbf{D}_k)}, \quad (32)$$

$$\delta_{s_k}(\mathbf{D}_k) \le \mu_{s_k-1}(\mathbf{D}_k). \quad (33)$$

Lemma 6 ([20] Lemma 4): Let $\mathbf{D}_k$'s be coordinate dictionaries such that $\delta_{s_k}(\mathbf{D}_k) < 1$. Then for any $\mathcal{J}_k \subset p_k$, $|\mathcal{J}_k| = s_k$, $\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}$ exists and

$$\left\|\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right\|_2 \le \frac{1}{1-\delta_{s_k}(\mathbf{D}_k)}, \quad \left\|\mathbf{D}_{k,\mathcal{J}_k}^{+}\right\|_2 \le \frac{1}{\sqrt{1-\delta_{s_k}(\mathbf{D}_k)}}, \quad (34)$$

and for any $\mathbf{D}_k'$ such that $\|\mathbf{D}_k - \mathbf{D}_k'\|_F \le \varepsilon_k < \sqrt{1-\delta_{s_k}(\mathbf{D}_k)}$:

$$1 - \delta_{s_k}(\mathbf{D}_k') \ge \left(\sqrt{1-\delta_{s_k}(\mathbf{D}_k)} - \varepsilon_k\right)^2 \triangleq 1 - \delta_k. \quad (35)$$

Lemma 7 ([20] Lemma 6): Given any $\mathbf{D}_k^1, \mathbf{D}_k^2 \in \mathcal{D}_k$, there exist $\mathbf{V}_k \in \mathbb{R}^{m_k \times p_k}$ with $\mathrm{diag}(\mathbf{D}_k^{1\top}\mathbf{V}_k) = \mathbf{0}$ and $\mathrm{diag}\left(\mathbf{V}_k^\top\mathbf{V}_k\right) = \mathbf{I}_{p_k}$ and a vector $\boldsymbol{\theta}_k \triangleq \boldsymbol{\theta}_k(\mathbf{D}_k^1, \mathbf{D}_k^2) \in [0,\pi]^{p_k}$, such that

$$\mathbf{D}_k^2 = \mathbf{D}_k^1\mathbf{C}_k(\boldsymbol{\theta}_k) + \mathbf{V}_k\mathbf{S}_k(\boldsymbol{\theta}_k), \quad (36)$$

where $\mathbf{C}_k(\boldsymbol{\theta}_k) \triangleq \mathrm{Diag}(\cos(\boldsymbol{\theta}_k))$ and $\mathbf{S}_k(\boldsymbol{\theta}_k) \triangleq \mathrm{Diag}\left(\sin(\boldsymbol{\theta}_k)\right)$. Moreover,

$$\frac{2}{\pi}\theta_{k,j} \le \|\mathbf{d}_{k,j}^2 - \mathbf{d}_{k,j}^1\|_2 = 2\sin\left(\frac{\theta_{k,j}}{2}\right) \le \theta_{k,j}, \text{and}$$

$$\frac{2}{\pi}\|\boldsymbol{\theta}_k\|_2 \le \|\mathbf{D}_k^2 - \mathbf{D}_k^1\|_F \le \|\boldsymbol{\theta}_k\|_2, \quad (37)$$

where $j \in [p_k]$. Similarly, there exists $\mathbf{V}_k'$ such that $\mathbf{D}_k^1 = \mathbf{D}_k^2\mathbf{C}_k(\boldsymbol{\theta}_k) + \mathbf{V}_k'\mathbf{S}_k(\boldsymbol{\theta}_k)$, where $\mathrm{diag}(\mathbf{D}_k^{2\top}\mathbf{V}_k') = \mathbf{0}$.

Lemma 8: Fix $\mathbf{D}_{1:K}$ and $\mathbf{D}_{1:K}^0$, and suppose $\{A_k\}$, $\{B_k\}$, $\{\delta_k\}$ satisfy the following:

$$\begin{aligned}
&A_k \ge \max\left\{\|\mathbf{D}_k^\top\mathbf{D}_k - \mathbf{I}_{p_k}\|_F, \|\mathbf{D}_k^{0\top}\mathbf{D}_k^0 - \mathbf{I}_{p_k}\|_F\right\},\\
&B_k \ge \max\left\{\|\mathbf{D}_k\|_2, \|\mathbf{D}_k^0\|_2\right\}, \text{and}\\
&\delta_k \ge \max\left\{\delta_{s_k}(\mathbf{D}_k), \delta_{s_k}(\mathbf{D}_k^0)\right\}.
\end{aligned} \quad (38)$$

Then for all $\boldsymbol{\theta}_k \triangleq \boldsymbol{\theta}_k(\mathbf{D}_k, \mathbf{D}_k^0)$, $k \in [K]$, we have

$$\begin{aligned}
&\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\\
&\ge \frac{s\mathbb{E}\{x^2\}}{2}\sum_{k\in[K]}\frac{\|\boldsymbol{\theta}_k\|_2}{p_k}\left[\|\boldsymbol{\theta}_k\|_2\left(1 - \frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k} - \bar{\lambda}\kappa_x^2\delta_{-k}\right)\right.\\
&\qquad \left. - \left(\delta_{-k} + 2\bar{\lambda}\prod_{i\in[K]}\frac{1}{1-\delta_i}\right)\bar{\lambda}\kappa_x^2\frac{s_k}{p_k}\frac{2A_kB_k}{1-\delta_k}\right],
\end{aligned} \quad (39)$$

where $\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\{|x|\}}$ and $\delta_{-k} \triangleq \prod_{\substack{i\in[K]\\ i\neq k}}\sqrt{\frac{1+\delta_i}{1-\delta_i}}$.

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

Proposition 2: Let the generating coordinate dictionaries $\{\mathbf{D}_k^0 \in \mathcal{D}_k\}$ satisfy:

$$\mu_s(\mathbf{D}^0) < \frac{1}{2}, \quad \max_k\{\delta_{s_k}(\mathbf{D}_k^0)\} < \frac{1}{4}. \quad (40)$$

Suppose $\bar{\lambda} = \frac{\lambda}{\mathbb{E}\{|x|\}} \le \frac{x_{\min}}{2\mathbb{E}\{|x|\}}$ and

$$\max_{k\in[K]}\{\varepsilon_k\} \le \min\left\{\bar{\lambda}C_{\max}, 0.15\right\}. \quad (41)$$

If the following is satisfied:

$$\frac{M_w}{M_x} < 3(1.5)^{K/2}\left(\bar{\lambda}KC_{\max} - \sum_{k\in[K]}\varepsilon_k\right), \quad (42)$$

then for any $\mathbf{D}_{1:K}$ such that $\mathbf{D}_k \in \mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)$, for $k \in [K]$, $\widehat{\mathbf{x}}$ that is defined in (28) is almost surely the minimizer of the map $\mathbf{x}' \mapsto \frac{1}{2}\|\mathbf{y} - (\bigotimes \mathbf{D}_k)\mathbf{x}'\|_2^2 + \lambda\|\mathbf{x}'\|_1$ and $\Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}) = \Delta f_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0)$.

Remark 3: Note that $\mu_s(\mathbf{D}^0) < \frac{1}{2}$ in (40) can be satisfied by ensuring that the right hand side of (6) is less than $\frac{1}{2}$. One way this can be ensured is by enforcing strict conditions on coordinate dictionaries; for instance, $\mu_{s_k}(\mathbf{D}_k^0) \leq \frac{1}{2^K}$.

The proof of Proposition 2 relies on the following lemmas and [20, Lemmas 10–13].

Lemma 9 ([20] Lemma 13): Assume $\mu_s(\mathbf{D}) < \frac{1}{2}$. If

$$\min_{j\in\mathcal{J}} |x_j| \geq 2\lambda, \text{ and } \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 < \lambda(1 - 2\mu_s(\mathbf{D})) \quad (43)$$

hold for generating $\mathbf{x}$, then $\widehat{\mathbf{x}}$ defined in (28) is the unique solution of $\min_{\mathbf{x}'} \frac{1}{2}\|\mathbf{y} - (\bigotimes \mathbf{D}_k)\,\mathbf{x}'\|_2 + \lambda\|\mathbf{x}'\|_1$.

Lemma 10: For any $\mathbf{D}^0 = \bigotimes \mathbf{D}_k^0$ and $\mathbf{D} = \bigotimes \mathbf{D}_k$ such that $\mathbf{D}_k \in \mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0)$, for $k \in [K]$, suppose the following inequalities are satisfied:

$$\max_{k\in[K]}\{\delta_{s_k}(\mathbf{D}_k^0)\} \leq \frac{1}{4}, \quad \text{and} \quad \max_{k\in[K]} \varepsilon_k \leq 0.15. \quad (44)$$

Then, we have

$$\mu_s(\mathbf{D}) \leq \mu_s(\mathbf{D}^0) + 2(1.5)^{K/2}\sqrt{s}\left(\sum_{k\in[K]} \varepsilon_k\right). \quad (45)$$

Proof of Theorem 1: To prove this theorem, we use Proposition 1 to show that $\Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}) > 0$, and then use Proposition 2 to show that $\Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}) = \Delta f_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0)$. The assumptions in (11) ensure that the conditions in (21) and (40) are satisfied for Proposition 1 and Proposition 2, respectively. Assumptions (12) and (14) ensure that the conditions in (22) and (24) are satisfied for Proposition 1, $\bar{\lambda} \leq \frac{x_{\min}}{2\mathbb{E}\{|x|\}}$ holds for Proposition 2, and $\max_{k\in[K]}\{C_{k,\min}\} < C_{\max}$. Hence, according to Proposition 1, $\Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}) > 0$ for all $\varepsilon_k \in (\bar{\lambda}C_{k,\min}, 0.15]$, $k \in [K]$. Finally, using the assumption in (16) implies $\Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}) = \Delta f_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0)$ for all $\varepsilon_k \leq \bar{\lambda}C_{\max}$, $k \in [K]$. Furthermore, the assumption in (14) implies $C_{\max}\bar{\lambda} \leq 0.15$. Consequently, for any $\{\varepsilon_k > 0, k \in [K]\}$ satisfying the conditions in (15), $\mathbf{D}_{1:K} \to f_{\mathbb{P}}(\mathbf{D}_{1:K})$ admits a local minimum $\widehat{\mathbf{D}} = \bigotimes \widehat{\mathbf{D}}_k$ such that $\widehat{\mathbf{D}}_k \in \mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0)$, $k \in [K]$.

IV. Finite Sample Identifiability Results

We now focus on leveraging Theorem 1 and solving (9) to derive finite-sample bounds for KS dictionary identifiability. Compared to Gribonval *et al.* [20], who use Lipschitz continuity of the objective function with respect to the larger KS dictionary, our analysis is based on "coordinate-wise Lipschitz continuity" with respect to the coordinate dictionaries.

Theorem 2: Suppose the observations are generated according to (7) and the dictionary coefficients follow the separable sparsity model of Section II such that (11) to (16) are satisfied. Next, fix any $\xi \in (0, \infty)$. Then, for any number of observations satisfying

$$N = \max_{k\in[K]} \Omega\Bigg(\frac{p_k^2(\xi + m_k p_k)}{(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda}))^2} \left(\frac{2^K(1+\bar{\lambda}^2)M_x^2}{s^2\mathbb{E}\{x^2\}^2} + \left(\frac{M_w}{s\mathbb{E}\{x^2\}}\right)^2\right)\Bigg), \quad (46)$$

with probability at least $1 - e^{-\xi}$, $\mathbf{D}_{1:K} \mapsto F_{\mathbf{Y}}(\mathbf{D}_{1:K})$ admits a local minimum $\widehat{\mathbf{D}} = \bigotimes \widehat{\mathbf{D}}_k$ such that $\widehat{\mathbf{D}}_k \in \mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0)$, for $k \in [K]$.

A. Discussion

Let us make some remarks about implications of Theorem 2. First, sample complexity has an inverse relationship with signal to noise ratio (SNR),[9] which we define as

$$\text{SNR} \triangleq \frac{\mathbb{E}\{\|\mathbf{x}\|_2^2\}}{\mathbb{E}\{\|\mathbf{w}\|_2^2\}} = \frac{s\mathbb{E}\{x^2\}}{m\mathbb{E}\{w^2\}}. \quad (47)$$

Looking at the terms on the right hand side of (46) in Theorem 2, $M_x/(s\mathbb{E}\{x^2\})$ is related to the deviation of $\|\mathbf{x}\|_2$ from its mean, $\mathbb{E}\{\|\mathbf{x}\|_2\}$, and depends on the coefficient distribution, while $M_w/(s\mathbb{E}\{x^2\})$ is related to $1/\text{SNR}$ and depends on the noise and coefficient distributions.

Second, we notice dependency of sample complexity on the recovery error of coordinate dictionaries. We can interpret ε_k as the recovery error for $\mathbf{D}_k^0$. Then, the sample complexity scaling in (46) is proportional to $\max_k \varepsilon_k^{-2}$. We note that the sample complexity results obtained in [20] that are independent of $\varepsilon \triangleq \|\mathbf{D} - \mathbf{D}^0\|_F$ only hold for the noiseless setting and the dependency on ε^{-2} is inevitable for noisy observations [20]. Furthermore, given the condition on the range of ε_k's in (15), ε_k's cannot be arbitrarily small, and will not cause N to grow arbitrarily large.

Third, we observe a linear dependence between the sample complexity scaling in (46) and coordinate dictionaries' dimensions, i.e., $\max_k \mathcal{O}(m_k p_k^3)$. Comparing this to the $\mathcal{O}(mp^3) = \mathcal{O}\left(\prod_k m_k p_k^3\right)$ scaling in the unstructured DL problem [20], the sample complexity in the KS-DL problem scales with the dimensions of the largest coordinate dictionary, as opposed to the dimensions of the larger KS dictionary.

We also compare this sample complexity upper bound scaling to the sample complexity lower bound scaling in our previous work [22, Corollary 1], where we obtained $N = \Omega\left(p\sum_k m_k p_k \varepsilon^{-2}/K\right)$ as a *necessary condition for recovery of KS dictionaries*.[10] In terms of overall error ε, our result translates into $N = \max_k \Omega\left\{2^K K^2 p(m_k p_k^3)\varepsilon^{-2}\right\}$ as a *sufficient* condition for recovery of coordinate dictionaries. The lower bound depended on the average dimension of the coordinate dictionaries, $\sum_k m_k p_k/K$, whereas we observe here a dependence on the dimensions of the coordinate dictionaries in terms of the maximum dimension, $\max_k m_k p_k$. We also observe an increase of order $\max_k p_k^2$ in the sample complexity upper bound scaling. This gap suggests that tighter bounds can be obtained for lower and/or upper bounds. A summary of these results is provided in Table I for a fixed K.

[9] Sufficient conditioning on N implies $\mathcal{O}$-scaling for sample complexity.

[10] We have the following relation between ε and ε_k's:

$$\varepsilon \leq \sum_{k\in[K]} \left(\prod_{\substack{i\in[K]\\ i\neq k}} \|\widetilde{\mathbf{D}}_k\|_F\right) \|\mathbf{D}_k - \mathbf{D}_k^0\|_F \leq \sqrt{p}\sum_{k\in[K]} \varepsilon_k.$$

Assuming all ε_k's are equal, this then implies $\varepsilon_k^2 \geq \varepsilon^2/(K^2 p)$.

TABLE I
COMPARISON OF UPPER AND LOWER BOUNDS ON THE SAMPLE COMPLEXITY OF DICTIONARY LEARNING FOR VECTORIZED DL AND KS DL

	Vectorized DL	KS DL
Minimax Lower Bound	$\frac{mp^2}{\varepsilon^2}$ [21]	$\frac{p\sum_k m_k p_k}{\varepsilon^2}$ [24]
Achievability Bound	$\frac{mp^3}{\varepsilon^2}$ [20]	$\max_k \frac{m_k p_k^3}{\varepsilon_k^2}$

B. Proof Outline

We follow a similar approach used in [20, Th. 2] for vectorized data. We show that, with high probability,

$$\Delta F_{\mathbf{Y}}(\varepsilon_{1:K}) \triangleq \inf_{\mathbf{D}_k \in \mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)} \Delta F_{\mathbf{Y}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0\right) \tag{48}$$

converges uniformly to its expectation,

$$\Delta f_{\mathbb{P}}(\varepsilon_{1:K}) \triangleq \inf_{\mathbf{D}_k \in \mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)} \Delta f_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0\right). \tag{49}$$

In other words, with high probability,

$$|\Delta F_{\mathbf{Y}}(\varepsilon_{1:K}) - \Delta f_{\mathbb{P}}(\varepsilon_{1:K})| \leq \eta_N, \tag{50}$$

where η_N is a parameter that depends on the probability and other parameters in the problem. This implies $\Delta F_{\mathbf{Y}}(\varepsilon_{1:K}) \geq \Delta f_{\mathbb{P}}(\varepsilon_{1:K}) - 2\eta_N$. In Theorem 1, we obtained conditions that ensure $\Delta f_{\mathbb{P}}(\varepsilon_{1:K}) > 0$. Thus, if $2\eta_N < \Delta f_{\mathbb{P}}(\varepsilon_{1:K})$ is satisfied, this implies $\Delta F_{\mathbf{Y}}(\varepsilon_{1:K}) > 0$, and we can use arguments similar to the proof of Theorem 1 to show that $\mathbf{D}_{1:K} \mapsto F_{\mathbf{Y}}(\mathbf{D}_{1:K})$ admits a local minimum $\widehat{\mathbf{D}} = \bigotimes \widehat{\mathbf{D}}_k$, such that $\widehat{\mathbf{D}}_k \in \mathbf{B}_{\varepsilon_k}(\mathbf{D}_k^0)$, for $k \in [K]$.

In Theorem 1, we showed that under certain conditions, $f_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0) = \Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0 | \boldsymbol{\sigma}\right)$. To find η_N, we uniformly bound deviations of $\mathbf{D}_{1:K} \mapsto \Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0 | \boldsymbol{\sigma}\right)$ from its expectation on $\left\{\mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)\right\}_{k=1}^K$. Our analysis is based on the *coordinate-wise Lipschitz continuity* property of $\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0 | \boldsymbol{\sigma}\right)$ with respect to coordinate dictionaries. Then, to ensure $2\eta_N < \Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0 | \boldsymbol{\sigma}\right)$, we show that $2\eta_N$ is less than the right-hand side of (23) and obtain conditions on the sufficient number of samples based on each coordinate dictionary dimension and recovery error.

The proof of Theorem 2 relies on the following definition and lemmas. The proofs of these are provided in Appendix B.

Definition 1 (Coordinate-wise Lipschitz continuity): A function $f : \mathcal{D}_1 \times \cdots \times \mathcal{D}_K \to \mathbb{R}$ is coordinate-wise Lipschitz continuous with constants $(L_1, \ldots, L_K)$ if there exist real constants $\{L_k \geq 0\}_{k=1}^K$, such that for $\{\mathbf{D}_k, \mathbf{D}_k' \in \mathcal{D}_k\}_{k=1}^K$:

$$|f(\mathbf{D}_{1:K}) - f(\mathbf{D}_{1:K}')| \leq \sum_{k\in[K]} L_k \|\mathbf{D}_k - \mathbf{D}_k'\|_F. \tag{51}$$

Lemma 11 (Rademacher averages [20]): Consider $\mathcal{F}$ to be a set of measurable functions on measurable set $\mathcal{X}$ and N i.i.d. random variables $X_1, \ldots, X_N \in \mathcal{X}$. Fix any $\xi \in (0, \infty)$. Assuming all functions are bounded by B, i.e., $|f(X)| \leq B$, almost surely, with probability at least $1 - e^{-\xi}$:

$$\sup_{f\in\mathcal{F}} \left(\frac{1}{N} \sum_{n\in[N]} f(X_n) - \mathbb{E}_X \{f(X)\} \right) \leq 2\sqrt{\frac{\pi}{2}} \mathbb{E}_{X,\beta_{1:N}} \left\{ \sup_{f\in\mathcal{F}} \left(\frac{1}{N} \sum_{n\in[N]} \beta_n f(X_n) \right) \right\} + B\sqrt{\frac{2\xi}{N}}, \tag{52}$$

where $\beta_{1:N}$'s are independent standard Gaussian random variables.

Lemma 12: Let $\mathcal{H}$ be a set of real-valued functions on $\mathbf{D}_k \in \overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0), k \in [K]$, that are bounded by B almost everywhere and are coordinate-wise Lipschitz continuous with constants $(L_1, \ldots, L_K)$. Let $h_1, h_2, \ldots, h_N$ be independent realizations from $\mathcal{H}$ with uniform Haar measure on $\mathcal{H}$. Then, fixing $\xi \in (0, \infty)$, we have with probability greater than $1 - e^{-\xi}$ that:

$$\sup_{\substack{\mathbf{D}_k \in \overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0) \\ k\in[K]}} \left| \frac{1}{N} \sum_{n\in[N]} h_n(\mathbf{D}_{1:K}) - \mathbb{E}\{h(\mathbf{D}_{1:K})\} \right| \leq 4\sqrt{\frac{\pi}{2N}} \left(\sum_{k\in[K]} L_k \varepsilon_k \sqrt{K m_k p_k} \right) + B\sqrt{\frac{2\xi}{N}}. \tag{53}$$

Lemma 13 ([20] Lemma 5): For any $\delta_k < 1$, $\mathbf{D}_k, \mathbf{D}_k'$ such that $\max(\delta_{s_k}(\mathbf{D}_k), \delta_{s_k}(\mathbf{D}_k')) \leq \delta_k$, and $\mathcal{J}_k \subset p_k, |\mathcal{J}_k| = s_k$, we have

$$\begin{aligned}
&\|\mathbf{I} - \mathbf{D}_{k,\mathcal{J}_k}^{+} \mathbf{D}_{k,\mathcal{J}_k}'\|_2 \leq (1-\delta_k)^{-1/2} \|\mathbf{D}_k - \mathbf{D}_k'\|_F, \\
&\|\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}} - \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}'}\|_2 \leq 2(1-\delta_k)^{-3/2} \|\mathbf{D}_k - \mathbf{D}_k'\|_F, \\
&\|\mathbf{D}_{k,\mathcal{J}_k}^{+} - \mathbf{D}'^{+}_{k,\mathcal{J}_k}\|_2 \leq 2(1-\delta_k)^{-1} \|\mathbf{D}_k - \mathbf{D}_k'\|_F, \text{and} \\
&\|\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}'}\|_2 \leq 2(1-\delta_k)^{-1/2} \|\mathbf{D}_k - \mathbf{D}_k'\|_F.
\end{aligned} \tag{54}$$

Lemma 14: Consider $\mathbf{D}_k^0 \in \mathcal{D}_k$ and ε_k's such that $\varepsilon_k < \sqrt{1 - \delta_{s_k}(\mathbf{D}_k^0)}$, for $k \in [K]$ and define $\sqrt{1-\delta_k} \triangleq \sqrt{1 - \delta_{s_k}(\mathbf{D}_k^0)} - \varepsilon_k > 0$. The function $\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0 | \boldsymbol{\sigma}\right)$ is almost surely coordinate-wise Lipschitz continuous on $\left\{\mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0)\right\}_{k=1}^K$ with Lipschitz constants

$$L_k \triangleq (1-\delta_k)^{-1/2} \left(M_x \left(\prod_{k\in[K]} \sqrt{1 + \delta_{s_k}(\mathbf{D}_k^0)} \right) + M_w + \lambda\sqrt{s} \prod_{k\in[K]} (1-\delta_k)^{-1/2} \right)^2, \tag{55}$$

and $\left|\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0 | \boldsymbol{\sigma}\right)\right|$ is almost surely bounded on $\left\{\mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0)\right\}_{k=1}^K$ by $\sum_{k\in[K]} L_k \varepsilon_k$.

Proof of Theorem 2: From Lemmas 12 and 14, we have that with probability at least $1-e^{-\xi}$:

$$\sup_{\substack{\mathbf{D}_k \in \overline{B}_{\varepsilon_k}(\mathbf{D}_k^0) \\ k\in[K]}} \left|\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) - \Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\right| \leq \sqrt{\frac{2}{N}} \sum_{k\in[K]} L_k \varepsilon_k \left(2\sqrt{\pi m_k p_k} + \sqrt{\xi}\right), \tag{56}$$

where L_k is defined in (55). From (56), we obtain $\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) > \Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) - 2\eta_N$ where $\eta_N = \sqrt{\frac{2}{N}}\sum_{k\in[K]} L_k\varepsilon_k\left(2\sqrt{\pi m_k p_k}+\sqrt{\xi}\right)$. In Theorem 1, we derived conditions that ensure $\Delta f_{\mathbf{y}}(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0) = \Delta\phi_{\mathbf{y}}(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma})$ and $\Delta f_{\mathbb{P}}(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0) = \Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)$. Therefore, given that the conditions in Theorem 1 are satisfied, $\Delta F_{\mathbf{Y}}(\varepsilon_{1:K}) > \Delta f_{\mathbb{P}}(\varepsilon_{1:K}) - 2\eta_N$, and the existence of a local minimum of $F_{\mathbf{Y}}(\mathbf{D}_{1:K})$ within radii ε_k around $\mathbf{D}_k^0$, $k\in[K]$, is guaranteed with probability at least $1-e^{-\xi}$ as soon as $2\eta_N < \Delta f_{\mathbb{P}}(\varepsilon_{1:K})$. According to (23), $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) \geq \frac{s\mathbb{E}\{x^2\}}{8}\sum_{k\in[K]}\frac{\varepsilon_k}{p_k}\left(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda})\right)$; therefore, it is sufficient to have for all $k\in[K]$:

$$\sqrt{\frac{8}{N}} L_k\varepsilon_k\left(2\sqrt{\pi m_k p_k}+\sqrt{\xi}\right) < \frac{s\mathbb{E}\{x^2\}\varepsilon_k\left(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda})\right)}{8p_k},$$

which translates into $N \geq \max_{k\in[K]} N_k$, where

$$N_k = \left(2\sqrt{\pi m_k p_k}+\sqrt{\xi}\right)^2 \left(\frac{2^{4.5}L_k p_k}{s\mathbb{E}\{x^2\}(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda}))}\right)^2. \tag{57}$$

Furthermore, we can upper bound L_k by

$$\begin{aligned} L_k &\overset{(a)}{\leq} \sqrt{2}\left(1.25^{K/2}M_x + M_w + 2^{K/2}\lambda\sqrt{s}\right)^2 \\ &\overset{(b)}{\leq} \sqrt{2}c_1\left(\left(1.25^K + 2^K\bar{\lambda}^2\right)M_x^2 + M_w^2\right), \end{aligned} \tag{58}$$

where c_1 is some positive constant, (a) follows from the fact that given the assumption in (21), assumptions in Lemma 14 are satisfied with $\sqrt{1-\delta_k} \geq \sqrt{1/2}$ for any $\varepsilon_k \leq 0.15$, and (b) follows from the following inequality:

$$\lambda = \bar{\lambda}\mathbb{E}\{|x|\} = \frac{1}{s}\bar{\lambda}\mathbb{E}\{\|\mathbf{x}\|_1\} \leq \frac{1}{\sqrt{s}}\bar{\lambda}\mathbb{E}\{\|\mathbf{x}\|_2\} \leq \frac{1}{\sqrt{s}}\bar{\lambda}M_x.$$

Substituting (58) in (57) and using $\left(\sqrt{\xi}+2\sqrt{\pi m_k p_k}\right)^2 \leq c_2(\xi + m_k p_k)$ for some positive constant c_2, we get

$$\begin{aligned} N_k &= \Omega\left(p_k^2(m_k p_k+\xi)\left(\frac{2^K(1+\bar{\lambda}^2)M_x^2 + M_w^2}{s^2\mathbb{E}\{x^2\}^2(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda}))^2}\right)\right) \\ &= \Omega\left(\frac{p_k^2(m_k p_k+\xi)}{(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda}))^2}\left(\frac{2^K(1+\bar{\lambda}^2)M_x^2}{s^2\mathbb{E}\{x^2\}^2} + \frac{M_w^2}{s^2\mathbb{E}\{x^2\}^2}\right)\right). \end{aligned}$$

and $N \geq \max_{k\in[K]} N_k$.

Remark 4: To bound deviations of $\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)$ from its mean, we can also use the bound provided in [29, Th. 1] that prove uniform convergence results using covering number arguments for various classes of dictionaries. In this case, we get $\eta_N \leq c\sqrt{\frac{\left(\sum_k m_k p_k + \xi\right)\log N}{N}}$ for some constant c, where an extra $\sqrt{\log N}$ term appears compared to (53). Therefore, Lemma 12 provides a tighter upper bound.

V. Conclusion

In this paper, we focused on local recovery of coordinate dictionaries comprising a Kronecker-structured dictionary used to represent Kth-order tensor data. We derived a sample complexity upper bound for coordinate dictionary identification up to specified errors by expanding the objective function with respect to individual coordinate dictionaries and using the coordinate-wise Lipschitz continuity property of the objective function. This analysis is local in the sense that it only guarantees existence of a local minimum of the KS-DL objective function within some neighborhood of true coordinate dictionaries. Global analysis of the KS-DL problem is left for future work. Our results hold for dictionary coefficients generated according to the separable sparsity model. This model has some limitations compared to the random sparsity model and we leave the analysis for the random sparsity model for future work also. Another future direction of possible interest includes providing practical KS-DL algorithms that achieve the sample complexity scaling of Theorem 2.

Appendix A

Proof of Lemma 2: To prove the existence of such a formation for any $K \geq 2$, we use induction. For $K=2$, we have

$$\begin{aligned} &(\mathbf{D}_1\otimes\mathbf{D}_2) - \left(\mathbf{D}_1^0\otimes\mathbf{D}_2^0\right) \\ &\quad= \left(\mathbf{D}_1 - \mathbf{D}_1^0\right)\otimes\mathbf{D}_2^0 + \mathbf{D}_1\otimes\left(\mathbf{D}_2 - \mathbf{D}_2^0\right) \\ &\quad= \left(\mathbf{D}_1 - \mathbf{D}_1^0\right)\otimes\mathbf{D}_2 + \mathbf{D}_1^0\otimes\left(\mathbf{D}_2 - \mathbf{D}_2^0\right). \end{aligned} \tag{59}$$

For K such that $K>2$, we assume the following holds:

$$\begin{aligned} &\bigotimes_{k\in[K]}\mathbf{D}_k - \bigotimes_{k\in[K]}\mathbf{D}_k^0 \\ &\quad= \sum_{k\in[K]}\widetilde{\mathbf{D}}_{k,1}\otimes\cdots\otimes\left(\mathbf{D}_k - \mathbf{D}_k^0\right)\otimes\cdots\otimes\widetilde{\mathbf{D}}_{k,K}. \end{aligned} \tag{60}$$

Then, for $K+1$, we have:

$$\begin{aligned} &\bigotimes_{k\in[K+1]}\mathbf{D}_k - \bigotimes_{k\in[K+1]}\mathbf{D}_k^0 \\ &= \left(\bigotimes_{k\in[K]}\mathbf{D}_k\right)\otimes\mathbf{D}_{K+1} - \left(\bigotimes_{k\in[K]}\mathbf{D}_k^0\right)\otimes\mathbf{D}_{K+1}^0 \\ &\overset{(a)}{=} \left(\bigotimes_{k\in[K]}\mathbf{D}_k - \bigotimes_{k\in[K]}\mathbf{D}_k^0\right)\otimes\mathbf{D}_{K+1}^0 \\ &\quad+ \left(\bigotimes_{k\in[K]}\mathbf{D}_k\right)\left(\mathbf{D}_{K+1} - \mathbf{D}_{K+1}^0\right) \\ &\overset{(b)}{=} \left(\sum_{k\in[K]}\widetilde{\mathbf{D}}_{k,1}\otimes\cdots\otimes\left(\mathbf{D}_k - \mathbf{D}_k^0\right)\otimes\cdots\otimes\widetilde{\mathbf{D}}_{k,K}\right) \\ &\qquad\otimes\mathbf{D}_{K+1}^0 + \left(\bigotimes_{k\in[K]}\mathbf{D}_k\right)\left(\mathbf{D}_{K+1} - \mathbf{D}_{K+1}^0\right) \\ &\overset{(c)}{=} \sum_{k\in[K+1]}\widetilde{\mathbf{D}}_{k,1}\otimes\cdots\otimes\left(\mathbf{D}_k - \mathbf{D}_k^0\right)\otimes\cdots\otimes\widetilde{\mathbf{D}}_{k,K+1}, \end{aligned} \tag{61}$$

where (a) follows from (59), (b) follows from (60) and (c) follows from replacing $\mathbf{D}^0_{K+1}$ with $\widetilde{\mathbf{D}}_{k,K+1}$ in the first K terms of the summation and $\mathbf{D}_k$'s with $\widetilde{\mathbf{D}}_{K+1,k}$, for $k \in [K]$, in the $(K+1)$th term of the summation.

Proof of Lemma 3: Using the same definition as Gribonval *et al.* [20, Definition 1], taking the derivative of $\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right)$ with respect to $\mathbf{x}$ and setting it to zero, we get the expression in (28) for $\widehat{\mathbf{x}}$. Substituting $\widehat{\mathbf{x}}$ in (27), we get

$$\begin{aligned}\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right) &= \frac{1}{2}\Bigg[\|\mathbf{y}\|_2^2 - \left(\left(\bigotimes \mathbf{D}^\top_{k,\mathcal{J}_k}\right)\mathbf{y} - \lambda\boldsymbol{\sigma}_{\mathcal{J}}\right)^\top \\ &\quad \left(\bigotimes (\mathbf{D}^\top_{k,\mathcal{J}_k}\mathbf{D}_{k,\mathcal{J}_k})^{-1}\right)\left(\left(\bigotimes \mathbf{D}^\top_{k,\mathcal{J}_k}\right)\mathbf{y} - \lambda\boldsymbol{\sigma}_{\mathcal{J}}\right)\Bigg] \\ &\overset{(a)}{=} \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\mathbf{y}^\top\left(\bigotimes \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\mathbf{y} \\ &\quad + \lambda\boldsymbol{\sigma}^\top_{\mathcal{J}}\left(\bigotimes \mathbf{D}^+_{k,\mathcal{J}_k}\right)\mathbf{y} - \frac{\lambda^2}{2}\boldsymbol{\sigma}^\top_{\mathcal{J}}\left(\bigotimes \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\boldsymbol{\sigma}_{\mathcal{J}},\end{aligned}$$

where (a) follows from (25).

Proof of Lemma 4: We use the expression for $\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right)$ from (29). For any $\mathbf{D} = \bigotimes \mathbf{D}_k, \mathbf{D}' = \bigotimes \mathbf{D}'_k$, $\mathbf{D}_k, \mathbf{D}'_k \in \mathcal{D}_k$, we have

$$\begin{aligned}\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) &= \phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right) - \phi_{\mathbf{y}}\left(\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) \\ &= \frac{1}{2}\mathbf{y}^\top\left(\bigotimes \mathbf{P}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \bigotimes \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\mathbf{y} \\ &\quad - \lambda\boldsymbol{\sigma}^\top_{\mathcal{J}}\left(\bigotimes \mathbf{D}'^{+}_{k,\mathcal{J}_k} - \bigotimes \mathbf{D}^+_{k,\mathcal{J}_k}\right)\mathbf{y} \\ &\quad + \frac{\lambda^2}{2}\boldsymbol{\sigma}^\top_{\mathcal{J}}\left(\bigotimes \mathbf{H}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \bigotimes \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\boldsymbol{\sigma}_{\mathcal{J}}. \qquad (62)\end{aligned}$$

We substitute $\mathbf{y} = (\bigotimes \mathbf{D}^0_k)\mathbf{x} + \mathbf{w} = (\bigotimes \mathbf{D}^0_{k,\mathcal{J}_k})\mathbf{x}_{\mathcal{J}} + \mathbf{w}$ and break up the sum in (62) into 6 terms:

$$\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) = \sum_{i\in[6]}\Delta\phi_i\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right), \qquad (63)$$

where

$$\begin{aligned}\Delta\phi_1\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) &= \frac{1}{2}\mathbf{x}^\top\left(\bigotimes \mathbf{D}^0_k\right)^\top \\ &\quad \left(\bigotimes \mathbf{P}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \bigotimes \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\left(\bigotimes \mathbf{D}^0_k\right)\mathbf{x} \\ &\overset{(a)}{=} \frac{1}{2}\mathbf{x}^\top\left(\bigotimes \mathbf{D}^0_k\right)^\top\Bigg(\sum_{k\in[K]}\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\otimes\cdots\otimes \\ &\quad \left(\mathbf{P}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\otimes\cdots\otimes\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\Bigg)\left(\bigotimes \mathbf{D}^0_k\right)\mathbf{x} \\ &= \frac{1}{2}\mathbf{x}^\top\Bigg(\sum_{k\in[K]}\left(\mathbf{D}^{0\top}_1\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\mathbf{D}^0_1\right)\otimes\cdots\otimes \\ &\quad \left(\mathbf{D}^{0\top}_k(\mathbf{P}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}})\mathbf{D}^0_k\right)\otimes\cdots\otimes \\ &\quad \left(\mathbf{D}^{0\top}_K\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\mathbf{D}^0_K\right)\Bigg)\mathbf{x},\end{aligned}$$

$$\begin{aligned}\Delta\phi_2\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) &= \mathbf{w}^\top\Bigg(\sum_{k\in[K]}\left(\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\mathbf{D}^0_1\right)\otimes\cdots\otimes \\ &\quad \left((\mathbf{P}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}})\mathbf{D}^0_k\right)\otimes\cdots\otimes\left(\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\mathbf{D}^0_K\right)\Bigg)\mathbf{x},\end{aligned}$$

$$\begin{aligned}\Delta\phi_3\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) &= \frac{1}{2}\mathbf{w}^\top\Bigg(\sum_{k\in[K]}\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\otimes\cdots\otimes \\ &\quad \left(\mathbf{P}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\otimes\cdots\otimes\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\Bigg)\mathbf{w},\end{aligned}$$

$$\begin{aligned}\Delta\phi_4\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) &= -\lambda\boldsymbol{\sigma}^\top_{\mathcal{J}}\Bigg(\sum_{k\in[K]}\left(\widetilde{\mathbf{D}}^+_{1,\mathcal{J}_1}\mathbf{D}^0_1\right)\otimes\cdots\otimes \\ &\quad \left((\mathbf{D}'^{+}_{k,\mathcal{J}_k} - \mathbf{D}^+_{k,\mathcal{J}_k})\mathbf{D}^0_k\right)\otimes\cdots\otimes\left(\widetilde{\mathbf{D}}^+_{K,\mathcal{J}_K}\mathbf{D}^0_K\right)\Bigg)\mathbf{x},\end{aligned}$$

$$\begin{aligned}\Delta\phi_5\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) &= -\lambda\boldsymbol{\sigma}^\top_{\mathcal{J}}\Bigg(\sum_{k\in[K]}\widetilde{\mathbf{D}}^+_{1,\mathcal{J}_1}\otimes\cdots\otimes \\ &\quad \left(\mathbf{D}'^{+}_{k,\mathcal{J}_k} - \mathbf{D}^+_{k,\mathcal{J}_k}\right)\otimes\cdots\otimes\widetilde{\mathbf{D}}^+_{K,\mathcal{J}_K}\Bigg)\mathbf{w}, \text{ and}\end{aligned}$$

$$\begin{aligned}\Delta\phi_6\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) &= \frac{\lambda^2}{2}\boldsymbol{\sigma}^\top_{\mathcal{J}}\Bigg(\sum_{k\in[K]}\mathbf{H}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\otimes\cdots\otimes \\ &\quad \left(\mathbf{H}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\otimes\cdots\otimes\mathbf{H}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\Bigg)\boldsymbol{\sigma}_{\mathcal{J}}, \qquad (64)\end{aligned}$$

where (a) follows from Lemma 2 and analysis for derivation of $\{\Delta\phi_i\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right)\}_{i=2}^6$ are omitted due to space constraints. Now, we set $\mathbf{D}' = \mathbf{D}^0$ and take the expectation of $\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\{\mathbf{D}^0_k\}|\boldsymbol{\sigma}\right)$ with respect to $\mathbf{x}$ and $\mathbf{w}$. Since the coefficient and noise vectors are uncorrelated,

$$\mathbb{E}\left\{\Delta\phi_2\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\} = \mathbb{E}\left\{\Delta\phi_5\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\} = 0.$$

We can restate the other terms as:

$$\begin{aligned}&\Delta\phi_1\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) \\ &\overset{(b)}{=} \frac{1}{2}\operatorname{Tr}\Bigg[\mathbf{x}_{\mathcal{J}}\mathbf{x}^\top_{\mathcal{J}}\sum_{k\in[K]}\left(\mathbf{D}^{0\top}_1\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\mathbf{D}^0_1\right)\otimes\cdots\otimes \\ &\quad \left(\mathbf{D}^{0\top}_k(\mathbf{I}_{m_k} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}})\mathbf{D}^0_k\right)\otimes\cdots\otimes\left(\mathbf{D}^{0\top}_K\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\mathbf{D}^0_K\right)\Bigg], \\ &\Delta\phi_3\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) \\ &= \frac{1}{2}\operatorname{Tr}\Bigg[\mathbf{w}\mathbf{w}^\top\Bigg(\sum_{k\in[K]}\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\otimes\cdots\otimes\left(\mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right) \\ &\quad \otimes\cdots\otimes\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\Bigg)\Bigg],\end{aligned}$$

$$\Delta\phi_4\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)$$
$$\overset{(c)}{=} -\lambda \operatorname{Tr}\Bigg[\mathbf{x}_{\mathcal{J}}\boldsymbol{\sigma}_{\mathcal{J}}^\top\Bigg(\sum_{k\in[K]}\left(\widetilde{\mathbf{D}}^+_{1,\mathcal{J}_1}\mathbf{D}^0_1\right)\otimes\cdots\otimes$$
$$\left(\mathbf{I}_{s_k}-\mathbf{D}^+_{k,\mathcal{J}_k}\mathbf{D}^0_k\right)\otimes\cdots\otimes\left(\widetilde{\mathbf{D}}^+_{K,\mathcal{J}_K}\mathbf{D}^0_K\right)\Bigg)\Bigg], \text{ and}$$
$$\Delta\phi_6\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)$$
$$=\frac{\lambda^2}{2}\operatorname{Tr}\Bigg[\boldsymbol{\sigma}_{\mathcal{J}}\boldsymbol{\sigma}_{\mathcal{J}}^\top\Bigg(\sum_{k\in[K]}\mathbf{H}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\otimes\cdots\otimes$$
$$\left(\mathbf{H}_{\mathbf{D}^0_{k,\mathcal{J}_k}}-\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\otimes\cdots\otimes\mathbf{H}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\Bigg)\Bigg], \tag{65}$$

where (b) and (c) follow from the facts that $\mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}}\mathbf{D}^0_k=\mathbf{D}^0_k$ and $\mathbf{D}^{0+}_{k,\mathcal{J}_k}\mathbf{D}^0_k=\mathbf{I}_{s_k}$, respectively. Taking the expectation of the terms in (65), we get

$$\mathbb{E}\left\{\Delta\phi_1\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}$$
$$\overset{(d)}{=}\frac{\mathbb{E}\{x^2\}}{2}\mathbb{E}_{\mathcal{J}}\Bigg\{\sum_{k\in[K]}\operatorname{Tr}\left[\mathbf{D}_1^{0\top}\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\mathbf{D}^0_1\right]\cdots$$
$$\operatorname{Tr}\left[\mathbf{D}_k^{0\top}(\mathbf{I}_{m_k}-\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}})\mathbf{D}^0_k\right]\cdots\operatorname{Tr}\left[\mathbf{D}_K^{0\top}\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\mathbf{D}^0_K\right]\Bigg\}$$
$$=\frac{\mathbb{E}\{x^2\}}{2}\sum_{k\in[K]}\mathbb{E}_{\mathcal{J}_1}\left\{\operatorname{Tr}\left[\mathbf{D}_1^{0\top}\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\mathbf{D}^0_1\right]\right\}\cdots$$
$$\mathbb{E}_{\mathcal{J}_k}\left\{\operatorname{Tr}\left[\mathbf{D}_k^{0\top}(\mathbf{I}_{m_k}-\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}})\mathbf{D}^0_k\right]\right\}\cdots$$
$$\mathbb{E}_{\mathcal{J}_K}\left\{\operatorname{Tr}\left[\mathbf{D}_K^{0\top}\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\mathbf{D}^0_K\right]\right\},$$
$$\mathbb{E}\{\Delta\phi_3\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\}$$
$$=\frac{\mathbb{E}\{w^2\}}{2}\mathbb{E}_{\mathcal{J}}\Bigg\{\operatorname{Tr}\Bigg[\sum_{k\in[K]}\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\otimes\cdots\otimes$$
$$\left(\mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}}-\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\otimes\cdots\otimes\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\Bigg]\Bigg\}$$
$$=\frac{\mathbb{E}\{w^2\}}{2}\mathbb{E}_{\mathcal{J}}\Bigg\{\sum_{k\in[K]}\operatorname{Tr}\left[\mathbf{P}_{\widetilde{\mathbf{D}}_{1,J_1}}\right]\cdots\operatorname{Tr}\left[\mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}}-\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right]$$
$$\cdots\operatorname{Tr}\left[\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\right]\Bigg\}$$
$$\overset{(e)}{=}0,$$
$$\mathbb{E}\left\{\Delta\phi_4\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}$$
$$=-\lambda\mathbb{E}\{|x|\}\sum_{k\in[K]}\mathbb{E}_{\mathcal{J}_1}\left\{\operatorname{Tr}\left[\widetilde{\mathbf{D}}^+_{1,\mathcal{J}_1}\mathbf{D}^0_1\right]\right\}\cdots$$
$$\mathbb{E}_{\mathcal{J}_k}\left\{\operatorname{Tr}\left[\mathbf{I}_{s_k}-\mathbf{D}^+_{k,\mathcal{J}_k}\mathbf{D}^0_k\right]\right\}\cdots\mathbb{E}_{\mathcal{J}_K}\left\{\operatorname{Tr}\left[\widetilde{\mathbf{D}}^+_{K,\mathcal{J}_K}\mathbf{D}^0_K\right]\right\},$$
$$\mathbb{E}\left\{\Delta\phi_6\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}$$
$$=\frac{\lambda^2}{2}\sum_{k\in[K]}\mathbb{E}_{\mathcal{J}_1}\left\{\operatorname{Tr}\left[\mathbf{H}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\right]\right\}\cdots$$
$$\mathbb{E}_{\mathcal{J}_k}\left\{\operatorname{Tr}\left[\mathbf{H}_{\mathbf{D}^0_{k,\mathcal{J}_k}}-\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right]\right\}\cdots\mathbb{E}_{\mathcal{J}_K}\left\{\operatorname{Tr}\left[\mathbf{H}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\right]\right\}. \tag{66}$$

where (d) follows from the relation $\operatorname{Tr}(\mathbf{A}\otimes\mathbf{B})=\operatorname{Tr}[\mathbf{A}]\operatorname{Tr}[\mathbf{B}]$ [26] and (e) follows from the fact that $\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}$'s are orthogonal projections onto subspaces of dimension s_k and $\operatorname{Tr}[\mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}}-\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}]=s_k-s_k=0$. Adding the terms in (66), we obtain the expression in (31).

Proof of Lemma 5: Equation (32) follows from the definition of RIP and (33) follows from Gerschgorin's disk theorem [26], [30], [31].

Proof of Lemma 8: To lower bound $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)$, we bound each term in (31) separately. For the first term $\mathbb{E}\left\{\Delta\phi_1\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}$, we have

$$\mathbb{E}_{\mathcal{J}_k}\left\{\operatorname{Tr}\left[\mathbf{D}_k^{0\top}\mathbf{P}_{\widetilde{\mathbf{D}}_{k,\mathcal{J}_k}}\mathbf{D}^0_k\right]\right\}=\mathbb{E}_{\mathcal{J}_k}\left\{\left\|\mathbf{P}_{\widetilde{\mathbf{D}}_{k,\mathcal{J}_k}}\mathbf{D}^0_{k,\mathcal{J}_k}\right\|_F^2\right\}. \tag{67}$$

If $\widetilde{\mathbf{D}}_k=\mathbf{D}^0_k$, then

$$\mathbb{E}_{\mathcal{J}_k}\left\{\left\|\mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}}\mathbf{D}^0_{k,\mathcal{J}_k}\right\|_F^2\right\}\overset{(a)}{=}\frac{s_k}{p_k}\left\|\mathbf{D}^0_k\right\|_F^2=s_k, \tag{68}$$

where (a) follows from [20, Lemma 15]. If $\widetilde{\mathbf{D}}_k=\mathbf{D}_k$, then

$$\mathbb{E}_{\mathcal{J}_k}\left\{\left\|\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\mathbf{D}^0_{k,\mathcal{J}_k}\right\|_F^2\right\}\overset{(b)}{=}\mathbb{E}_{\mathcal{J}_k}\left\{\left\|[\mathbf{D}_k\mathbf{C}_k^{-1}]_{\mathcal{J}_k}\right\|_F^2\right\}$$
$$\overset{(c)}{=}\frac{s_k}{p_k}\left\|\mathbf{D}_k\mathbf{C}_k^{-1}\right\|_F^2\overset{(d)}{=}\frac{s_k}{p_k}\sum_{j=1}^{p_k}\frac{1}{\cos^2(\theta_{(k,j)})}\overset{(e)}{\geq}\frac{s_k}{p_k}p_k=s_k,$$

where (b) is a direct consequence of Lemma 7; we can write $\mathbf{D}^0_k=\mathbf{D}_k\mathbf{C}_k^{-1}-\mathbf{V}_k\mathbf{T}_k$ where $\mathbf{C}_k=\operatorname{Diag}(\cos(\boldsymbol{\theta}_k))$, $\mathbf{T}_k=\operatorname{Diag}(\tan(\boldsymbol{\theta}_k))$ and $\theta_{k,j}$ denotes the angle between $\mathbf{d}_{k,j}$ and $\mathbf{d}^0_{k,j}$. Hence $\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\mathbf{D}^0_{k,\mathcal{J}_k}=[\mathbf{D}_k\mathbf{C}_k^{-1}]_{\mathcal{J}_k}$. Moreover, (c) follows from [20, Lemma 15], (d) follows from the fact that $\|\mathbf{d}_{k,j}\|_2=1$, and (e) follows from the fact that $\cos(\theta_{k,j})<1$. Similarly, we have

$$\mathbb{E}_{\mathcal{J}_k}\left\{\operatorname{Tr}\left[\mathbf{D}_k^{0\top}(\mathbf{I}_{m_k}-\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}})\mathbf{D}^0_k\right]\right\}$$
$$=\mathbb{E}_{\mathcal{J}_k}\left\{\left\|(\mathbf{I}_{m_k}-\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}})\mathbf{D}^0_{k,\mathcal{J}_k}\right\|_F^2\right\}$$
$$\overset{(f)}{\geq}\frac{s_k}{p_k}\|\boldsymbol{\theta}_k\|_2^2\left(1-\frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k}\right), \tag{69}$$

where (f) follows from similar arguments as in Gribonval *et al.* [20, eq. (72)]. Putting it all together, we have

$$\begin{aligned}
&\mathbb{E}\left\{\Delta\phi_1\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}\\
&\geq \frac{\mathbb{E}\{x^2\}}{2}\sum_{k\in[K]}\left(\prod_{\substack{i\in[K]\\ i\neq k}} s_i\right)\frac{s_k}{p_k}\|\boldsymbol{\theta}_k\|_2^2\left(1-\frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k}\right)\\
&= \frac{s\mathbb{E}\{x^2\}}{2}\sum_{k\in[K]}\frac{\|\boldsymbol{\theta}_k\|_2^2}{p_k}\left(1-\frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k}\right). \qquad (70)
\end{aligned}$$

Next, to lower bound $\mathbb{E}\left\{\Delta\phi_4\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}$, we upper bound $\left|\mathbb{E}\left\{\Delta\phi_4\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}\right|$. If $\widetilde{\mathbf{D}}_k=\mathbf{D}^0_k$, we have

$$\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[{\mathbf{D}^0}^+_{k,\mathcal{J}_k}\mathbf{D}^0_{k,\mathcal{J}_k}\right]\right\}=\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{I}_{s_k}\right]\right\}=s_k, \qquad (71)$$

otherwise, if $\widetilde{\mathbf{D}}_k=\mathbf{D}_k$, we get

$$\begin{aligned}
&\left|\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[{\mathbf{D}_{k,\mathcal{J}_k}}^+\mathbf{D}^0_k\right]\right\}\right|\\
&\overset{(g)}{\leq} s_k\mathbb{E}_{\mathcal{J}_k}\left\{\left\|\mathbf{D}^+_{k,\mathcal{J}_k}\mathbf{D}^0_{k,\mathcal{J}_k}\right\|_2\right\}\\
&\leq s_k\mathbb{E}_{\mathcal{J}_k}\left\{\|\mathbf{D}^+_{k,\mathcal{J}_k}\|_2\|\mathbf{D}^0_{k,\mathcal{J}_k}\|_2\right\}\\
&\overset{(h)}{\leq} s_k\left(\frac{1}{\sqrt{1-\delta_{s_k}(\mathbf{D}_k)}}\right)\left(\sqrt{1+\delta_{s_k}(\mathbf{D}^0_k)}\right)\\
&\overset{(i)}{\leq} s_k\sqrt{\frac{1+\delta_k}{1-\delta_k}}, \qquad (72)
\end{aligned}$$

where (g) follows from the fact that for a square matrix $\mathbf{A}\in\mathbb{R}^{q\times q}$, $\mathrm{Tr}\left[\mathbf{A}\right]\leq q\|\mathbf{A}\|_2$, (f) follows from (32) and (34) and (i) follows from (38). Similar to [20, eq. (73)], we also have

$$\begin{aligned}
&\left|\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{I}_{s_k}-\mathbf{D}^+_{k,\mathcal{J}_k}\mathbf{D}^0_k\right]\right\}\right|\\
&\qquad\leq\frac{s_k}{p_k}\frac{\|\boldsymbol{\theta}_k\|_2^2}{2}+\frac{s_k^2}{p_k^2}\frac{A_kB_k}{1-\delta_k}\|\boldsymbol{\theta}_k\|_2. \qquad (73)
\end{aligned}$$

Thus, defining $\delta_{-k}\triangleq\prod_{\substack{i\in[K]\\ i\neq k}}\sqrt{\frac{1+\delta_i}{1-\delta_i}}$, we get

$$\begin{aligned}
&\mathbb{E}\left\{\Delta\phi_4\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}\\
&\geq-\lambda\mathbb{E}\{|x|\}\sum_{k\in[K]}\delta_{-k}\left(\prod_{\substack{i\in[K]\\ i\neq k}}s_i\right)\\
&\qquad\left(\frac{s_k}{p_k}\frac{\|\boldsymbol{\theta}_k\|_2^2}{2}+\frac{s_k^2}{p_k^2}\frac{A_kB_k}{1-\delta_k}\|\boldsymbol{\theta}_k\|_2\right)\\
&=-\lambda s\mathbb{E}\{|x|\}\sum_{k\in[K]}\frac{\delta_{-k}}{p_k}\left(\frac{\|\boldsymbol{\theta}_k\|_2^2}{2}+\frac{s_k}{p_k}\frac{A_kB_k}{1-\delta_k}\|\boldsymbol{\theta}_k\|_2\right). \qquad (74)
\end{aligned}$$

To lower bound $\mathbb{E}\left\{\Delta\phi_6\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}$, we upper bound $\left|\mathbb{E}\left\{\Delta\phi_6\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}\right|$. For any $\widetilde{\mathbf{D}}_k$, we have

$$\left|\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{H}_{\widetilde{\mathbf{D}}_{k,\mathcal{J}_k}}\right]\right\}\right|\leq\mathbb{E}_{\mathcal{J}_k}\left\{s_k\left\|\mathbf{H}_{\widetilde{\mathbf{D}}_{k,\mathcal{J}_k}}\right\|_2\right\}\overset{(j)}{\leq}\frac{s_k}{1-\delta_k}, \qquad (75)$$

where (j) follows from (34) and (38). Similar to Gribonval *et al.* [20, eq. (74)], we also have

$$\left|\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{H}_{\mathbf{D}^0_{k,\mathcal{J}_k}}-\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right]\right\}\right|\leq\frac{s_k^2}{p_k^2}\frac{4A_kB_k}{(1-\delta_k)^2}\|\boldsymbol{\theta}_k\|_2.$$

Thus, we get

$$\begin{aligned}
&\mathbb{E}\{\Delta\phi_6\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\}\\
&\geq-\frac{\lambda^2}{2}\sum_{k\in[K]}\left(\prod_{\substack{i\in[K]\\ i\neq k}}\frac{s_i}{1-\delta_i}\right)\left(\frac{s_k^2}{p_k^2}\frac{4A_kB_k}{(1-\delta_k)^2}\|\boldsymbol{\theta}_k\|_2\right)\\
&=-\frac{\lambda^2 s}{2}\sum_{k\in[K]}\frac{1}{p_k}\left(\prod_{i\in[K]}\frac{1}{1-\delta_i}\right)\left(\frac{s_k}{p_k}\frac{4A_kB_k}{1-\delta_k}\|\boldsymbol{\theta}_k\|_2\right). \qquad (76)
\end{aligned}$$

Adding (70), (74), and (76), we get (39).

Proof of Proposition 1: To show that $\Delta\phi_{\mathbb{P}}(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma})>0$, we use Lemma 8 and prove that the right hand side of (39) is positive under certain conditions. First, we ensure the conditions in (35) and (38) hold for Lemma 6 and Lemma 8, respectively. We set $\delta_k=\frac{1}{2}$, $\delta_{s_k}(\mathbf{D}_k)=\frac{1}{2}$ and $\delta_{s_k}(\mathbf{D}^0_k)=\frac{1}{4}$, for $k\in[K]$. For $\varepsilon_k\leq 0.15$, this ensures:

$$\begin{aligned}
&\sqrt{1-\delta_{s_k}(\mathbf{D}_k)}\geq\sqrt{1-\delta_{s_k}(\mathbf{D}^0_k)}-\varepsilon_k,\ \text{and}\\
&\max\left\{\delta_{s_k}(\mathbf{D}^0_k),\delta_{s_k}(\mathbf{D}_k)\right\}\leq\delta_k, \qquad (77)
\end{aligned}$$

and implies $\delta_k<1$ (condition for Lemmas 4 and 13). Next, we find conditions that guarantee:

$$\frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k}+\bar{\lambda}\kappa_x^2\delta_{-k}\overset{(a)}{=}\frac{2B_k^2s_k}{p_k}+\bar{\lambda}\kappa_x^2\left(3\right)^{(K-1)/2}\leq\frac{1}{2}, \qquad (78)$$

where (a) follows from replacing δ_k with $\frac{1}{2}$. If we take $\frac{s_k}{p_k}\leq\frac{1}{8B_k^2}$ and $\bar{\lambda}\leq\frac{1}{8\times3^{(K-1)/2}}$, given the fact that $\kappa_x^2\leq1$, (78) is satisfied.[11] Consequently, we can restate (39) as

$$\begin{aligned}
\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\geq\frac{s\mathbb{E}\{x^2\}}{4}\sum_{k\in[K]}\frac{\|\boldsymbol{\theta}_k\|_2}{p_k}\Bigg[\|\boldsymbol{\theta}_k\|_2\\
-8\left(3^{(K-1)/2}+2^{(K+1)}\bar{\lambda}\right)\bar{\lambda}\kappa_x^2\frac{s_k}{p_k}A_kB_k\Bigg]. \qquad (79)
\end{aligned}$$

From [20, Proof of Proposition 2], we use the following relations:

$$B_k\leq B^0_k+\varepsilon_k\leq B^0_k+1,\quad A_k\leq A^0_k+2B_k\varepsilon_k,\quad k\in[K], \qquad (80)$$

[11] These numbers are chosen for a simplified proof and can be modified.

where $A_k^0 \triangleq \|\mathbf{D}_k^{0\top}\mathbf{D}_k^0 - \mathbf{I}_{p_k}\|_F$ and $B_k^0 \triangleq \|\mathbf{D}_k^0\|_2$ and (80) follows from matrix norm inequalities [20]. Defining $\gamma_k \triangleq 16(3^{(K-1)/2} + 2^{(K+1)}\bar{\lambda})\bar{\lambda}\kappa_x^2\frac{B_k^2 s_k}{p_k}$ for $k \in [K]$ and using $\kappa_x^2 \leq 1$, we have

$$\begin{aligned}\gamma_k &\leq 2\left(3^{(K-1)/2} + \frac{2^{(K+1)}}{8\times 3^{(K-1)/2}}\right)\left(\frac{1}{8\times 3^{(K-1)/2}}\right)\\ &\leq 2\left(\frac{1}{8}+\frac{4}{64}\right) \leq \frac{1}{2}.\end{aligned} \tag{81}$$

Then, for $\mathbf{D}_k \in \mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)$, $k \in [K]$, we get

$$\begin{aligned}&\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\\ &\overset{(b)}{\geq} \frac{s\mathbb{E}\{x^2\}}{4}\sum_{k\in[K]}\frac{\varepsilon_k}{p_k}\left(\varepsilon_k - \frac{\gamma_k}{2}\frac{A_k}{B_k}\right)\\ &\overset{(c)}{\geq} \frac{s\mathbb{E}\{x^2\}}{4}\sum_{k\in[K]}\frac{\varepsilon_k}{p_k}\left(\varepsilon_k - \frac{\gamma_k}{2}\frac{A_k^0+2B_k\varepsilon_k}{B_k}\right)\\ &\geq \frac{s\mathbb{E}\{x^2\}}{4}\sum_{k\in[K]}\frac{\varepsilon_k}{p_k}\left(\varepsilon_k(1-\gamma_k) - \frac{\gamma_k}{2}\frac{A_k^0}{B_k}\right)\\ &\overset{(d)}{\geq} \frac{s\mathbb{E}\{x^2\}}{8}\sum_{k\in[K]}\frac{\varepsilon_k}{p_k}\left(\varepsilon_k - \gamma_k\frac{A_k^0}{B_k}\right),\end{aligned} \tag{82}$$

where (b) follows from (79), (c) follows from (80), and (d) follows from (81). Hence, we can write

$$\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) \geq \frac{s\mathbb{E}\{x^2\}}{8}\sum_{k\in[K]}\frac{\varepsilon_k}{p_k}\left(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda})\right), \tag{83}$$

where we define

$$\begin{aligned}\varepsilon_{k,\min}(\bar{\lambda}) &\triangleq \gamma_k\frac{A_k^0}{B_k}\\ &= 16\left(3^{(K-1)/2}+2^{(K+1)}\bar{\lambda}\right)\bar{\lambda}\kappa_x^2\frac{s_k}{p_k}A_k^0B_k\\ &= \frac{2}{3^{(K+1)/2}}\left(3^{(K-1)/2}+2^{(K+1)}\bar{\lambda}\right)\bar{\lambda}C_{k,\min},\end{aligned} \tag{84}$$

and $C_{k,\min}$ is defined in (13). The lower bound in (83) holds for any $\varepsilon_k \leq 0.15$ and $\mathbf{D}_k \in \mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)$, $k \in [K]$. Finally, since $3^{(K-1)/2} + 2^{(K+1)}\bar{\lambda} \leq 0.5 \times 3^{(K+1)/2}$, the assumption $\bar{\lambda} \leq 0.15/(\max_{k\in[K]} C_{k,\min})$ implies that $\varepsilon_{k,\min}(\bar{\lambda}) \leq 0.15$ for $k \in [K]$. Therefore, $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) > 0$ for all $\varepsilon_k \in (\varepsilon_{k,\min}(\bar{\lambda}), 0.15]$, $k \in [K]$. ■

Proof of Lemma 10: Considering $j \notin \mathcal{J}$, associated with $(j_1,\ldots,j_k) \notin (\mathcal{J}_1\times\cdots\times\mathcal{J}_K)$, we have

$$\begin{aligned}&\|\mathbf{D}_{\mathcal{J}}^\top\mathbf{d}_j\|_1\\ &\overset{(a)}{\leq} \|\mathbf{D}_{\mathcal{J}}^{0\top}\mathbf{d}_j^0\|_1 + \|\mathbf{D}_{\mathcal{J}}^{0\top}(\mathbf{d}_j-\mathbf{d}_j^0)\|_1 + \|(\mathbf{D}_{\mathcal{J}}-\mathbf{D}_{\mathcal{J}}^0)^\top\mathbf{d}_j\|_1\\ &\leq \mu_s(\mathbf{D}^0) + \sqrt{s}\left[\|\mathbf{D}_{\mathcal{J}}^{0\top}(\mathbf{d}_j-\mathbf{d}_j^0)\|_2 + \|(\mathbf{D}_{\mathcal{J}}-\mathbf{D}_{\mathcal{J}}^0)^\top\mathbf{d}_j\|_2\right]\\ &\leq \mu_s(\mathbf{D}^0) + \sqrt{s}\left[\left\|\bigotimes\mathbf{D}_{k,\mathcal{J}_k}^{0\top}\right\|_2\left\|\bigotimes(\mathbf{d}_{k,j_k}-\mathbf{d}_{k,j_k}^0)\right\|_2\right.\\ &\quad\left.+\left\|\bigotimes\mathbf{D}_{k,\mathcal{J}_k}-\bigotimes\mathbf{D}_{k,\mathcal{J}_k}^0\right\|_2\|\mathbf{d}_j\|_2\right]\\ &\overset{(b)}{\leq} \mu_s(\mathbf{D}^0) + \sqrt{s}\left[\left(\prod_{k\in[K]}\sqrt{1+\delta_{s_k}(\mathbf{D}_k^0)}\right)\right.\\ &\quad\left(\sum_{k\in[K]}\left\|\widetilde{\mathbf{d}}_{1,j_1}\right\|_2\cdots\|\mathbf{d}_{k,j_k}-\mathbf{d}_{k,j_k}^0\|_2\cdots\left\|\widetilde{\mathbf{d}}_{k,j_K}\right\|_2\right)\\ &\quad\left.+\sum_{k\in[K]}\left\|\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}\right\|_2\cdots\|\mathbf{D}_{k,\mathcal{J}_k}-\mathbf{D}_{k,\mathcal{J}_k}^0\|_2\cdots\left\|\widetilde{\mathbf{D}}_{k,\mathcal{J}_k}\right\|_2\right]\\ &\overset{(c)}{\leq} \mu_s(\mathbf{D}^0) + \sqrt{s}\left[\left(\prod_{k\in[K]}\sqrt{1+\delta_{s_k}(\mathbf{D}_k^0)}\right)\left(\sum_{k\in[K]}\varepsilon_k\right)\right.\\ &\quad\left.+\sum_{k\in[K]}\left(\prod_{\substack{i\in[K]\\ i\neq k}}\left\|\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}\right\|_2\right)\varepsilon_k\right]\\ &\overset{(d)}{\leq} \mu_s(\mathbf{D}^0) + 2(1.5)^{K/2}\sqrt{s}\left(\sum_{k\in[K]}\varepsilon_k\right),\end{aligned} \tag{85}$$

where (a) follows from the triangle inequality, (b) follows from (26), (c) follows from (33), and, (d) follows from substituting the upper bound value from (44) for $\delta_{s_k}(\mathbf{D}_k^0)$. For $\widetilde{\mathbf{D}}_i = \mathbf{D}_i^0$, $\left\|\mathbf{D}_{i,\mathcal{J}_i}^0\right\|_2 \leq \sqrt{1+\delta_{s_i}(\mathbf{D}_i^0)} \leq \sqrt{\frac{5}{4}} < 1.5$ and for $\widetilde{\mathbf{D}}_i = \mathbf{D}_i$, according to (80), we have $\|\mathbf{D}_{i,\mathcal{J}_i}\|_2 \leq \left\|\mathbf{D}_{i,\mathcal{J}_i}^0\right\|_2 + \varepsilon_i \leq \sqrt{\frac{5}{4}} + 0.15 < 1.5$. ■

Proof of Proposition 2: We follow a similar approach to Gribonval *et al.* [20]. We show that the conditions in (43) hold for Lemma 9. We have

$$\begin{aligned}&\left\|\mathbf{y}-\left(\bigotimes\mathbf{D}_k\right)\mathbf{x}\right\|_2\\ &\leq \left\|\left(\bigotimes\mathbf{D}_{k,\mathcal{J}_k}^0-\bigotimes\mathbf{D}_{k,\mathcal{J}_k}\right)\mathbf{x}_{\mathcal{J}}\right\|_2 + \|\mathbf{w}\|_2\\ &\leq M_x\sum_{k\in[K]}\|\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}\otimes\cdots\otimes(\mathbf{D}_{k,\mathcal{J}_k}^0-\mathbf{D}_{k,\mathcal{J}_k})\otimes\cdots\otimes\\ &\quad\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}\|_2 + M_w\\ &\leq M_x\sum_{k\in[K]}\left\|\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}\right\|_2\cdots\|\mathbf{D}_{k,\mathcal{J}_k}^0-\mathbf{D}_{k,\mathcal{J}_k}\|_2\cdots\left\|\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}\right\|_2\\ &\quad+M_w\\ &\leq M_x\sum_{k\in[K]}\left(\prod_{\substack{i\in[K]\\ i\neq k}}\left\|\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}\right\|_2\right)\varepsilon_k + M_w\\ &\overset{(a)}{\leq} (1.5)^{(K-1)/2}M_x\sum_{k\in[K]}\varepsilon_k + M_w,\end{aligned} \tag{86}$$

where (a) follows from (40) and the fact that for $\widetilde{\mathbf{D}}_i = \mathbf{D}_i^0$, $\left\|\mathbf{D}_{i,\mathcal{J}_i}^0\right\|_2 \leq \sqrt{1+\delta_{s_i}(\mathbf{D}_i^0)} \leq \sqrt{\frac{5}{4}} < 1.5$ and for $\widetilde{\mathbf{D}}_i = \mathbf{D}_i$, according to (80), we have $\|\mathbf{D}_{i,\mathcal{J}_i}\|_2 \leq \left\|\mathbf{D}_{i,\mathcal{J}_i}^0\right\|_2 + \varepsilon_i \leq \sqrt{\frac{5}{4}} + 0.15 < 1.5$. Hence, we get

$$\begin{aligned}
&\lambda(1-2\mu_s(\mathbf{D})) - \left\|\mathbf{y} - \left(\bigotimes \mathbf{D}_k\right)\mathbf{x}\right\|_2 \\
&\geq \lambda(1-2\mu_s(\mathbf{D})) - (1.5)^{(K-1)/2} M_x \sum_{k\in[K]} \varepsilon_k - M_w \\
&\overset{(b)}{\geq} \lambda(1-2\mu_s(\mathbf{D}^0)) - (1.5)^{K/2}\left(4\lambda\sqrt{s} + (1.5)^{-1/2} M_x\right) \\
&\quad \sum_{k\in[K]} \varepsilon_k - M_w \\
&\overset{(c)}{\geq} \lambda(1-2\mu_s(\mathbf{D}^0)) - 3(1.5)^{K/2} M_x \sum_{k\in[K]} \varepsilon_k - M_w \\
&= 3(1.5)^{K/2} M_x \left(K\bar{\lambda} C_{\max} - \sum_{k\in[K]} \varepsilon_k\right) - M_w, \qquad (87)
\end{aligned}$$

where (b) follows from (45) and (c) follows from (43) ($2\lambda\sqrt{s} \leq x_{\min}\sqrt{s} \leq M_x$) and (45). If $\varepsilon_k < C_{\max}\bar{\lambda}$, $k \in [K]$, the assumption on the noise level in (42) implies that the right-hand side of (87) is greater than zero and $\lambda(1-2\mu_s(\mathbf{D})) > \|\mathbf{y} - (\bigotimes \mathbf{D}_k)\mathbf{x}\|_2$. Thus, according to Lemma 9, $\widehat{\mathbf{x}}$ is almost surely the unique solution of $\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - (\bigotimes \mathbf{D}_k)\mathbf{x}'\|_2 + \lambda\|\mathbf{x}'\|_1$ and $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}, \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) = \Delta f_{\mathbb{P}}\left(\mathbf{D}_{1:K}, \mathbf{D}_{1:K}^0\right)$. ■

Appendix B

Proof of Lemma 12: According to Lemma 11, we have to upper bound $\mathbb{E}\{\sup_{\mathbf{D}_k \in \overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0), k\in[K]} |\frac{1}{N}\sum_{n\in[N]} \beta_n h_n(\mathbf{D}_{1:K})|\}$. Conditioned on the draw of functions $h_1, \ldots, h_N$, consider the Gaussian processes $A_{\mathbf{D}_{1:K}} = \frac{1}{N}\sum_{n\in[N]} \beta_n h_n(\mathbf{D}_{1:K})$ and $C_{\mathbf{D}_{1:K}} = \sqrt{\frac{K}{N}}\sum_{k\in[K]}(L_k \sum_{i\in[m_k]}\sum_{j\in[p_k]} \zeta_{ij}^k (\mathbf{D}_k - \mathbf{D}_k^0)_{ij})$, where $\{\beta_n\}_{n=1}^N$'s and $\{\zeta_{ij}^k\}, k \in [K], i \in [m_k], j \in [p_k]$'s are independent standard Gaussian vectors. We have

$$\begin{aligned}
&\mathbb{E}\left\{\left|A_{\mathbf{D}_{1:K}} - A_{\mathbf{D}'_{1:K}}\right|^2\right\} \\
&= \frac{1}{N^2}\left|\sum_{n\in[N]} h_n(\mathbf{D}_{1:K}) - h_n(\mathbf{D}'_{1:K})\right|^2 \\
&\overset{(a)}{\leq} \frac{1}{N}\left(\sum_{k\in[K]} L_k\|\mathbf{D}_k - \mathbf{D}'_k\|_F\right)^2 \\
&\overset{(b)}{\leq} \frac{K}{N}\sum_{k\in[K]} L_k^2\|\mathbf{D}_k - \mathbf{D}'_k\|_F^2 \\
&= \mathbb{E}\left\{\left|C_{\mathbf{D}_{1:K}} - C_{\mathbf{D}'_{1:K}}\right|^2\right\}, \qquad (88)
\end{aligned}$$

where (a) follows from coordinate-wise Lipschitz continuity of h and (b) follows from Cauchy-Schwartz inequality. Hence, using Slepian's Lemma [32], we get

$$\begin{aligned}
\mathbb{E}\left\{\sup_{\substack{\mathbf{D}_k \in \overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0)\\ k\in[K]}} A_{\mathbf{D}_{1:K}}\right\} &\leq \mathbb{E}\left\{\sup_{\substack{\mathbf{D}_k \in \overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0)\\ k\in[K]}} C_{\mathbf{D}_{1:K}}\right\} \\
&= \sqrt{\frac{K}{N}}\left(\sum_{k\in[K]} L_k\varepsilon_k\mathbb{E}\{\|\zeta^k\|_F\}\right) \\
&= \sqrt{\frac{K}{N}}\left(\sum_{k\in[K]} L_k\varepsilon_k\sqrt{m_k p_k}\right). \qquad (89)
\end{aligned}$$

Thus, we obtain $\mathbb{E}\{\sup_{\substack{\mathbf{D}_k \in \overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0)\\ k\in[K]}} |\frac{1}{N}\sum_{n\in[N]} \beta_n h_n(\mathbf{D}_{1:K})|\} \leq 2\sqrt{\frac{K}{N}}(\sum_{k\in[K]} L_k\varepsilon_k\sqrt{m_k p_k})$.

Proof of Lemma 14: We expand $\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)$ according to (63) and bound each term of the sum separately. Looking at the first term, we get

$$\begin{aligned}
\left|\Delta\phi_1\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\right| &\overset{(a)}{=} \left|\frac{1}{2}\mathbf{x}^\top\mathbf{D}^{0\top}\left(\sum_{k\in[K]} \mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}} \otimes \cdots \otimes \right.\right. \\
&\quad \left.\left.\left(\mathbf{P}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right) \otimes \cdots \otimes \mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\right)\mathbf{D}^0\mathbf{x}\right| \\
&\overset{(b)}{\leq} \frac{1}{2}\|\mathbf{x}\|_2^2\left(\prod_{k\in[K]}\left\|\mathbf{D}_{k,\mathcal{J}_k}^0\right\|_2^2\right)\left(\sum_{k\in[K]}\left\|\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}^0} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right\|_2\right. \\
&\quad \left.\left(\prod_{\substack{i\in[K]\\ i\neq k}}\left\|\mathbf{P}_{\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}}\right\|_2\right)\right) \\
&\overset{(c)}{\leq} M_x^2\left(\prod_{k\in[K]}\left(1+\delta_{s_k}(\mathbf{D}_k^0)\right)\right) \\
&\quad \left(\sum_{k\in[K]}(1-\delta_k)^{-1/2}\|\mathbf{D}_k - \mathbf{D}_k^0\|_F\right), \qquad (90)
\end{aligned}$$

where (a) follows from (64), (b) follows from the fact that $\left\|\mathbf{D}_{\mathcal{J}}^0\right\|_2 = \prod_{k\in[K]}\left\|\mathbf{D}_{k,\mathcal{J}_k}^0\right\|_2$, and (c) follows from the definition of RIP, (54), and $\left\|\mathbf{P}_{\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}}\right\|_2 = 1$. Following a similar approach and expanding the rest of the terms, we get

$$\begin{aligned}
&\left|\Delta\phi_2\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\right| \\
&\leq \|\mathbf{w}\|_2\|\mathbf{x}\|_2\left(\prod_{k\in[K]}\left\|\mathbf{D}_{k,\mathcal{J}_k}^0\right\|_2^2\right) \\
&\quad \left(\sum_{k\in[K]}\left\|\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}^0} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right\|_2\left(\prod_{\substack{i\in[K]\\ i\neq k}}\left\|\mathbf{P}_{\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}}\right\|_2\right)\right) \\
&\overset{(d)}{\leq} 2M_w M_x\left(\prod_{k\in[K]}\left(1+\delta_{s_k}(\mathbf{D}_k^0)\right)^{1/2}\right)
\end{aligned}$$

$$\left(\sum_{k\in[K]}(1-\delta_k)^{-1/2}\|\mathbf{D}_k-\mathbf{D}_k^0\|_F\right),$$

$$\left|\Delta\phi_3\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\right| \le \frac{1}{2}\|\mathbf{w}\|_2^2 \left(\sum_{k\in[K]}\left\|\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}^0}-\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right\|_2\left(\prod_{\substack{i\in[K]\\ i\neq k}}\left\|\mathbf{P}_{\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}}\right\|_2\right)\right)$$

$$\le M_w^2\left(\sum_{k\in[K]}(1-\delta_k)^{-1/2}\|\mathbf{D}_k-\mathbf{D}_k^0\|_F\right),$$

$$\left|\Delta\phi_4\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\right| = \lambda\|\boldsymbol{\sigma}_{\mathcal{J}}\|_2\|\mathbf{x}\|_2\left(\prod_{k\in[K]}\left\|\mathbf{D}_{\mathcal{J}_k}^0\right\|_2\right) \left(\sum_{k\in[K]}\left\|\mathbf{D}^{0+}_{k,\mathcal{J}_k}-\mathbf{D}^{+}_{k,\mathcal{J}_k}\right\|_2\left(\prod_{\substack{i\in[K]\\ i\neq k}}\left\|\widetilde{\mathbf{D}}^{+}_{i,\mathcal{J}_i}\right\|_2\right)\right)$$

$$\overset{(e)}{\le} 2\lambda\sqrt{s}M_x\left(\prod_{k\in[K]}\left(1+\delta_{s_k}(\mathbf{D}_k^0)\right)^{1/2}\right) \left(\sum_{k\in[K]}(1-\delta_k)^{-1}\left(\prod_{\substack{i\in[K]\\ i\neq k}}(1-\delta_i)^{-1/2}\right)\|\mathbf{D}_k-\mathbf{D}_k^0\|_F\right),$$

$$\left|\Delta\phi_5\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\right| = \lambda\|\boldsymbol{\sigma}_{\mathcal{J}}\|_2\|\mathbf{w}\|_2 \left(\sum_{k\in[K]}\left\|\mathbf{D}^{0+}_{k,\mathcal{J}_k}-\mathbf{D}^{+}_{k,\mathcal{J}_k}\right\|_2\left(\prod_{\substack{i\in[K]\\ i\neq k}}\left\|\widetilde{\mathbf{D}}^{+}_{i,\mathcal{J}_i}\right\|_2\right)\right)$$

$$\le 2\lambda\sqrt{s}M_w \left(\sum_{k\in[K]}(1-\delta_k)^{-1}\left(\prod_{\substack{i\in[K]\\ i\neq k}}(1-\delta_i)^{-1/2}\right)\|\mathbf{D}_k-\mathbf{D}_k^0\|_F\right),$$

$$\left|\Delta\phi_6\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\right| = \frac{\lambda^2}{2}\|\boldsymbol{\sigma}_{\mathcal{J}}\|_2^2 \left(\sum_{k\in[K]}\left\|\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}^0}-\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right\|_2\left(\prod_{\substack{i\in[K]\\ i\neq k}}\left\|\mathbf{H}_{\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}}\right\|_2\right)\right)$$

$$\overset{(f)}{\le}\lambda^2 s\left(\sum_{k\in[K]}(1-\delta_k)^{-\frac{3}{2}}\left(\prod_{\substack{i\in[K]\\ i\neq k}}(1-\delta_i)^{-1}\right)\|\mathbf{D}_k-\mathbf{D}_k^0\|_F\right),$$

where (e) and (f) follow from (34) and (54). Adding all the terms together, we get

$$\left|\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\right| \le \sum_{k\in[K]}L_k\|\mathbf{D}_k-\mathbf{D}_k^0\|_F. \quad (91)$$

where L_k is defined in (55).

APPENDIX C

Proof of the coherence relation for KS dictionaries: To prove (6), we define the set $\mathcal{A}=\{\forall j_k\in\mathcal{J}_k,(j_1,\ldots,j_K)\notin(\mathcal{J}_1,\ldots,\mathcal{J}_K)\}$. We have

$$\begin{aligned}\mu_s(\mathbf{D}) &= \max_{|\mathcal{J}|\le s}\max_{j\notin\mathcal{J}}\|\mathbf{D}_{\mathcal{J}}^{\top}\mathbf{d}_j\|_1\\ &= \max_{\substack{|\mathcal{J}_k|\le s_k\\ k\in[K]}}\max_{\mathcal{A}}\left\|\left(\bigotimes\mathbf{D}_{k,\mathcal{J}_k}^{\top}\right)\left(\bigotimes\mathbf{d}_{k,j_k}\right)\right\|_1\\ &= \max_{\substack{|\mathcal{J}_k|\le s_k\\ k\in[K]}}\max_{\mathcal{A}}\left\|\bigotimes\mathbf{D}_{k,\mathcal{J}_k}^{\top}\mathbf{d}_{k,j_k}\right\|_1\\ &= \max_{\substack{|\mathcal{J}_k|\le s_k\\ k\in[K]}}\max_{\mathcal{A}}\prod_{k\in[K]}\left\|\mathbf{D}_{k,\mathcal{J}_k}^{\top}\mathbf{d}_{k,j_k}\right\|_1\\ &\le \max_{k\in[K]}\mu_{s_k}(\mathbf{D}_k)\left(\prod_{\substack{i\in[K],\\ i\neq k}}(1+\mu_{s_i-1}(\mathbf{D}_i))\right). \quad (92)\end{aligned}$$

REFERENCES

[1] Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "Identification of Kronecker-structured dictionaries: An asymptotic analysis," in *Proc. IEEE 7th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process.*, Dec. 2017, pp. 1–5.

[2] A. Smilde, R. Bro, and P. Geladi, *Multi-Way Analysis: Applications in the Chemical Sciences*. Hoboken, NJ, USA: Wiley, 2005.

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

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

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

[6] 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, Nov. 2006.

[7] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 494–504, Feb. 2012.

[8] C. F. Caiafa and A. Cichocki, "Multidimensional compressed sensing and their applications," *Wiley Interdisciplinary Rev., Data Mining Knowl. Discovery*, vol. 3, no. 6, pp. 355–380, Nov./Dec. 2013.

[9] S. Hawe, M. Seibert, and M. Kleinsteuber, "Separable dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 438–445.

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

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

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

[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 Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process.*, Dec. 2017, pp. 1–5.

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

[15] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries," in *Proc. 27th Annu. Conf. Learn. Theory*, 2014, vol. 35, pp. 1–15.

[16] 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, Jan. 2017.

[17] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Proc. 25th Annu. Conf. Learn. Theory*, 2014, vol. 35, pp. 1–28.
[18] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *Appl. Comput. Harmon. Anal.*, vol. 37, no. 3, pp. 464–491, Nov. 2014.
[19] K. Schnass, "Local identification of overcomplete dictionaries," *J. Mach. Learn. Res.*, vol. 16, pp. 1211–1242, Jun. 2015.
[20] 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, Nov. 2015.
[21] A. Jung, Y. C. Eldar, and N. Görtz, "On the minimax risk of dictionary learning," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1501–1515, Mar. 2015.
[22] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds for Kronecker-structured dictionary learning," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2016, pp. 1148–1152.
[23] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Sample complexity bounds for dictionary learning of tensor data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 4501–4505.
[24] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds on dictionary learning for tensor data," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2706–2726, Apr. 2018.
[25] C. F. Caiafa and A. Cichocki, "Computing sparse representations of multidimensional signals using Kronecker bases," *Neural Comput.*, vol. 25, no. 1, pp. 186–220, Jan. 2013.
[26] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1991.
[27] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9/10, pp. 589–592, 2008.
[28] S. Jokar and V. Mehrmann, "Sparse solutions to underdetermined Kronecker product systems," *Linear Algebra Appl.*, vol. 431, no. 12, pp. 2437–2447, Dec. 2009.
[29] 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, Jun. 2015.
[30] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[31] G. H. Golub and C. F. Van Loan, *Matrix Computations*, vol. 3. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 2012.
[32] P. Massart, *Concentration Inequalities and Model Selection*, vol. 6. New York, NY, USA: Springer, 2007.

Zahra Shakeri (S'15) received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2013, the M.Sc. degree in electrical and computer engineering, in 2016, from the Rutgers University, Piscataway, NJ, USA, where she is currently working toward the Ph.D. degree. Her research interests include machine learning, statistical signal processing, and multidimensional data processing. She is a member of the INSPIRE Laboratory.

Anand D. Sarwate (S'99–M'09–SM'14) received the B.S. degree 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 has been 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, USA. His research interests include information theory, machine learning, signal processing, optimization, and privacy and security. He was recipient of the A. Walter Tyson Assistant Professor Award from the Rutgers School of Engineering, 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 a recepient of the National Defense Science and Engineering Graduate Fellowship from 2002 to 2005. He is a member of Phi Beta Kappa and Eta Kappa Nu.

Waheed U. Bajwa (S'98–M'09–SM'13) received the B.E. (Hons.) degree in electrical engineering from the National University of Sciences and Technology, Islamabad, Pakistan, in 2001, and the M.S. and Ph.D. degrees in electrical engineering from the University of Wisconsin-Madison, Madison, WI, USA, in 2005 and 2009, respectively. From 2009 to 2010, he was a Postdoctoral Research Associate in the Program in Applied and Computational Mathematics, Princeton University, and from 2010 to 2011, a Research Scientist with the Department of Electrical and Computer Engineering, Duke University. Since 2011, he has been with Rutgers University, Piscataway, NJ, USA, where he is currently an Associate Professor with the Department of Electrical and Computer Engineering and an Associate Member of the graduate faculty with the Department of Statistics and Biostatistics. His research interests include statistical signal processing, high-dimensional statistics, machine learning, harmonic analysis, inverse problems, and networked systems.

He is a recepient of a number of awards in his career including the Best in Academics Gold Medal and President's 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 and 2017), and Rutgers University's Presidential Fellowship for Teaching Excellence (2017). He is a coinvestigator on the work that received the Cancer Institute of New Jersey's Gallo Award for Scientific Excellence in 2017, a coauthor 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 for the IEEE SIGNAL PROCESSING LETTERS (2014–2017), coguest edited a special issue of the Elsevier *Physical Communication Journal* on "Compressive Sensing in Communications" (2012), cochaired 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 serving as the Technical Co-Chair of the IEEE SPAWC 2018 Workshop, Senior Area Editor for the IEEE SIGNAL PROCESSING LETTERS, an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS, and serves on the MLSP, SAM, and SPCOM Technical Committees of the IEEE Signal Processing Society.