DICTIONARY LEARNING AND MULTIDIMENSIONAL PROCESSING FOR TENSOR DATA

by

ZAHRA SHAKERI

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Electrical and Computer Engineering

Written under the direction of

Waheed U. Bajwa

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

OCTOBER, 2019

ABSTRACT OF THE DISSERTATION

Dictionary Learning and Multidimensional Processing for Tensor Data

By ZAHRA SHAKERI

Dissertation Director:
Waheed U. Bajwa

Modern machine learning and signal processing relies on finding meaningful and succinct representations of data. While most works in the literature have focused on finding representations of vector data, many of today's data are collected using various sensors and have a multidimensional structure. This dissertation addresses the problem of feature learning for tensor (i.e., multiway) data, which are defined as data having multiple modes. The work presented in this dissertation aims to study the theoretical and algorithmic aspects of dictionary learning from tensor data and further investigate the computational aspects of exploiting the structure of tensor data in wireless communication systems. The dissertation has been divided into three main parts.

The first part of the dissertation is focused on the theoretical aspects of Kronecker-structured dictionary learning from tensor data. Here, the structure of tensor data is exploited by requiring that the dictionary underlying the vectorized versions of tensor data samples be Kronecker structured. That is, it is comprised of coordinate dictionaries that independently transform various modes of the tensor data. The presented results are primarily stated in terms of lower and upper bounds on the sample complexity of

dictionary learning, defined as the number of samples needed to reconstruct the true structured dictionary underlying the tensor data from noisy samples. These results highlight the effects of different parameters on the sample complexity of the problem and also bring out the potential advantages of structured dictionary learning from tensor data.

The second part of this dissertation focuses on extending the Kronecker-structured dictionary learning model to a less restrictive class of dictionaries referred to as low-separation-rank dictionary learning, while still exploiting the structure of tensor data in the underlying dictionary. Various computational algorithms are developed to learn such dictionaries in cases where tensor data are available in batch or are streaming in an online manner. Numerical experiments are provided to demonstrate the performance of the provided algorithms for synthetic tensor data representation and real-world image data denoising. These experiments highlight the advantages of the low-separation-rank dictionary learning model over Kronecker-structured dictionary learning for complex data classes such as images in the denoising problem.

The final part of the dissertation focuses on another application of sparse representations of tensor data and studies the sparse channel estimation problem in massive multiple-input multiple-output orthogonal frequency division multiplexing (MIMO-OFDM) systems. By modeling the underlying wireless channel as a tensor, a sparse tensor recovery technique is used to estimate the channel using lower computational resources and storage at the receiver compared to vectorized representation methods. Numerical experiments are provided to compare the performance of the estimation algorithms corresponding to vectorized and tensor formulations. These results also highlight the effects of various training signal parameters on the channel estimation performance.

Acknowledgements

I am profoundly grateful to my adviser, Prof. Waheed Bajwa, for his guidance, support and encouragement throughout the years that I have been a graduate student at Rutgers. I am especially thankful to him for helping me overcome my doubts, truly believe in myself, and have persistence and perseverance in research. He has been an extraordinarily perceptive and knowledgeable mentor and has taught me how to approach and solve research problems through many insightful discussions and helpful comments. In addition, I would also like to thank him and the Electrical and Computer Engineering department for providing me with financial support during my graduate studies at Rutgers.

I am also grateful to my other thesis committee members, Prof. Anand Sarwate, Prof. Athina Petropulu, and Dr. Brian Eriksson for their valuable insights on my dissertation and numerous helpful suggestions. I would specifically like to express my gratitude to Anand, who has been a mentor and collaborator in many of my projects. His passion and excitement for mathematical problems and his deep knowledge of theoretical research has been a constant inspiration to me and has helped me grow as a researcher. I am also grateful to Athina, whom I have deeply admired from the first semester I joined Rutgers. She has constantly encouraged and empowered female students in the Electrical and Engineering department and given us determination and confidence. I would also like to thank Brian for being a brilliant mentor while I was an intern at Technicolor. He truly helped me understand the difference between academic and industry research and bridge the gap between them.

Next, I would like to thank my lab members at INSPIRE laboratory for their support and making my studies more enjoyable at Rutgers. I would specifically like to thank Haroon, Talal, and Tong for many helpful discussions and fun times during my first

years as a graduate student. I would also like to thank my colleague Mohsen for being an excellent collaborator and a wonderful friend throughout our graduate studies.

Completing this degree could not be possible without the support of my amazing friends. I would like to express my gratitude to Parishad, Elham, Hanifa, Marjan, and Arman for encouraging me, listening to me and being there for me throughout these years. I would also like to thank all my Iranian friends at Rutgers, without whom I would have much less laughs and memories.

I would also like to thank my parents and my siblings, Amin and Elaheh, for their immense love and support. My mother and father have always encouraged to strive for the greatest and I am grateful to them for making many sacrifices to see me thrive and succeed.

Finally, I started this journey six years ago with my best friend and the love of my life, Alireza. I am appreciative for his constant love and understanding and I am truly grateful to him for being an enthusiastic cheerleader in all decisions that I have made and for helping me get through hard times.

Dedication

To my parents, for their endless love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Roughly speaking, data representation entails transforming raw data from its original domain to another domain in which it can be processed more effectively and efficiently. In particular, the performance of any information processing algorithm is dependent on the representation on which it was built on [2]. Data-driven representation approaches infer transforms from the data to yield efficient representations. Such techniques generally outperform model-based techniques that use predetermined bases to transform data. This success is attributed to the fact that the learned transformations in data-driven approaches are tuned to the input signals [3, 4].

Since contemporary data are often high dimensional and high volume, we need efficient algorithms to manage them. In addition, rapid advances in sensing and data acquisition technologies in recent years have resulted in individual data samples or signals with *multimodal* structures. Such data are often termed *tensors* or multiway arrays [5]. Examples of tensor data include hyperspectral images that have three modes (two spatial and one spectral), colored videos that have four modes (two spatial, one depth, and one temporal), and dynamic magnetic resonance imaging in a clinical trial that has five modes (three spatial, one temporal, and one subject).

In this thesis, we primarily focus on data-driven representations for tensor data. As data collection systems grow and proliferate, we will require efficient data representations for processing, storage, and retrieval of tensor data. Dictionary learning (DL) is a technique for finding sparse representations of data and has applications in various tasks such as image denoising and inpainting, audio processing, and classification [4, 6–8]. In traditional DL literature, tensor data are converted into one-dimensional data by vectorizing the signals. Recent works have shown that many multidimensional signals

can be decomposed into a superposition of separable atoms [9–11]. In this case, a sequence of independent transformations on different data dimensions can be carried out using Kronecker-structured (KS) matrices.

To provide some insights into the usefulness of KS dictionaries for tensor data, consider the 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 $p \geq 2^{25}$ is the number of dictionary columns. 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 $p_3 \geq 32$ are the number of columns of the coordinate dictionaries. 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, the fundamental problem of theoretical limits on the learning of KS dictionaries underlying Kth-order tensor data remains open.

Although KS-DL approaches may require lower sample and computational complexity and have better storage efficiency over unstructured DL [12], the KS-DL model makes a strong separability assumption among different modes of tensor data, namely, various modes of data can be transformed using independent transformations. This assumption is often too restrictive for many classes of data [13]. This results in an unfavorable tradeoff between model compactness and representation power. To overcome this limitation, a generalization of the KS-DL model referred to as *learning a mixture of separable dictionaries* or *low separation-rank DL* (LSR-DL) can be used that assumes the dictionary is comprised of summation of KS matrices. The LSR-DL model interpolates between the under-parameterized KS-DL model and the over-parameterized unstructured model.

Most of the focus of prior work on tensor data representation has been on the application of image data representation [14–16]. Another application of tensor data representation is the problem of channel estimation in multiple-input and multiple-output orthogonal frequency division multiplexing (MIMO-OFDM) systems that can be

modeled as a tensor with four modes (angle of arrival, angle of departure, delay spread, Doppler spread). In the case where the underlying multipath channel is approximately sparse in the angle-delay-Doppler domain, the channel estimation problem can be solved using a compressed sensing (CS) framework in which the channel response is sparsely represented in some predetermined bases [17]. Taking the tensor structure of the channel into account, one can take advantage of sparse tensor recovery techniques to estimate the channel using lower computational power and storage at the receiver compared to vectorized recovery techniques.

1.1 Thesis Statement

To address the problem of feature extraction for tensor data, we can assume a structure on the tensor of interest through tensor decompositions such as the CANDECOMP/PARAFAC (CP) decomposition [18], Tucker decomposition [19], PARATUCK decomposition [5], and Tensor-Train decomposition [20] 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 [11, 14, 16].

Hence, the thesis of this dissertation is: *"Taking the structure of tensor data into account in representation learning has fundamental advantages over vectorized learning techniques as it can lead to more compressed and efficient representations that can be obtained using less number of data samples compared to vectorized learning techniques. Furthermore, multidimensional processing techniques can be utilized to process tensor data that require less computational power and storage compared to vectorized processing methods."*

1.2 Major Contributions

In this thesis, we first aim to provide a theoretical understanding of the fundamental limits of DL methods that explicitly account for the multidimensional structure of data through KS dictionaries. We also provide structured DL algorithms for efficient tensor data representation. Finally, we investigate the computational advantages of using tensor recovery techniques over vectorized methods. In order to support our thesis, we have developed new theory and methods in the dissertation for some of the fundamental problems arising in finding sparse representations and processing of tensor data. Below, we highlight some of the primary aspects of these contributions:

Our first major contribution, which appears in **Chapter 3**, is focused on using an information-theoretic approach to provide lower bounds for the worst-case mean-squared error (MSE) of KS dictionaries that generate Kth-order tensor data. Furthermore, we also show that for a special case of $K = 2$, there exists an estimator whose MSE meets the derived lower bounds.

Our second major contribution, which appears in **Chapter 4**, examines 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 true KS dictionary that generate Kth-order tensor data.

These results suggest the sample complexity of KS-DL for tensor data can be significantly lower than that for unstructured data: for unstructured data, the sample complexity lower bound scales linearly with the product of the dictionary dimensions, whereas for tensor data the bound scales linearly with the sum of the product of the dimensions of the coordinate dictionaries comprising the KS dictionary. Furthermore, we show that the sample complexity upper bound in the KS-DL problem scales with the dimensions of the largest coordinate dictionary, as opposed to the dimensions of the larger KS dictionary when the multidimensional structure is ignored.

Our third major contribution, which appears in **Chapter 5**, is generalizing the KS-DL model to LSR-DL and developing various algorithms to learn LSR dictionaries in both batch and online settings. We also conduct numerical experiments to show the

effectiveness of the proposed model and the performance of the developed algorithms for synthetic data representation and real-world data image denoising.

Our final major contribution, which appears in **Chapter 6**, is focused on the problem of sparse channel estimation in massive MIMO-OFDM systems. Here, two formulations are investigated for training-based channel estimation. For the first formulation, theoretical guarantees for reliable channel recovery are provided based on the total number of parameters in the training signal. Moreover, a tensor recovery technique is used to estimate the sparse channel in the second formulation. By exploiting the tensor structure of the channel, computationally simpler sparse recovery algorithms are utilized to recover the channel in this formulation. Finally, numerical experiments are provided to investigate the channel estimation performance as a function of the used formulation and other training parameters.

1.3 Notational Convention and Definitions

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 $\mathbf{x}_k$. Let $\mathbf{X}_{\mathcal{I}}$ be the matrix consisting of columns of $\mathbf{X}$ with indices $\mathcal{I}$, $\mathbf{X}^{\mathcal{T}}$ be the matrix consisting of rows of $\mathbf{X}$ with indices $\mathcal{T}$ and $\mathbf{I}_d$ be the $d \times d$ identity matrix. 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 also use matrices indexed by multiple letters, such as $\mathbf{X}_{(a,b,c)}$, in which case its j-th column is denoted by $\mathbf{x}_{(a,b,c),j}$.

The function supp(.) denotes the locations of the nonzero entries of $\mathbf{X}$. We use $\text{vec}(\mathbf{X})$ 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 $\text{diag}\,(\mathbf{X})$ to denote the vector comprised of the diagonal elements of $\mathbf{X}$ and $\text{Diag}\,(\mathbf{v})$ 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 $\text{sign}(\mathbf{v})$, are equal to $\text{sign}(v_i) = v_i/|v_i|$, for $v_i \neq 0$, and $\text{sign}(v_i) = 0$ 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). We use $[K]$ to denote $\{1, 2, \ldots, K\}$ and $\mathbf{X}_{1:K}$ to denote $\{\mathbf{X}_k\}_{k=1}^K$.

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 $\|\mathbf{X}\|_2$, $\|\mathbf{X}\|_F$, and $\|\mathbf{X}\|_{\text{tr}}$ are the spectral, Frobenius, and trace (nuclear) norms of $\mathbf{X}$, respectively. Moreover, $\|\mathbf{X}\|_1 \triangleq \sum_{i,j} |x_{i,j}|$ denotes the sum of absolute values of entries of $\mathbf{X}$.

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

$$\begin{aligned}
\mathcal{S}_\varepsilon(\mathbf{X}^0) &\triangleq \left\{\mathbf{X} \in \mathcal{X} : \|\mathbf{X} - \mathbf{X}^0\|_F = \varepsilon\right\}, \\
\mathcal{B}_\varepsilon(\mathbf{X}^0) &\triangleq \left\{\mathbf{X} \in \mathcal{X} : \|\mathbf{X} - \mathbf{X}^0\|_F < \varepsilon\right\}, \\
\bar{\mathcal{B}}_\varepsilon(\mathbf{X}^0) &\triangleq \left\{\mathbf{X} \in \mathcal{X} : \|\mathbf{X} - \mathbf{X}^0\|_F \leq \varepsilon\right\}.
\end{aligned} \tag{1.1}$$

Note that while $\mathcal{S}_\varepsilon(\mathbf{X}^0)$ represents the surface of a sphere, we use the term "sphere" for simplicity. Furthermore, $P_{\mathcal{B}_1}(\mathbf{u})$ denotes the projection of $\mathbf{u}$ on the closed unit ball, i.e.,

$$P_{\mathcal{B}_1}(\mathbf{u}) = \begin{cases} \mathbf{u}, & \text{if } \|\mathbf{u}\|_2 \leq 1, \\ \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, & \text{otherwise.} \end{cases} \tag{1.2}$$

We use $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$ if for sufficiently large $n \in \mathbb{N}$, $f(n) < C_1 g(n)$ and $f(n) > C_2 g(n)$, respectively, for some positive constants C_1 and C_2.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 $\Delta f(\mathbf{X}; \mathbf{Y}) \triangleq f(\mathbf{X}) - f(\mathbf{Y})$.

A frame $\mathbf{F} \in \mathbb{R}^{m \times p}$, $m \leq p$, is defined as a collection of vectors $\{\mathbf{f}_i \in \mathbb{R}^m\}_{i=1}^p$ in some separable Hilbert space $\mathcal{H}$, that satisfy $c_1 \|\mathbf{v}\|_2^2 \leq \sum_{i=1}^p |\langle \mathbf{f}_i, \mathbf{v} \rangle|^2 \leq c_2 \|\mathbf{v}\|_2^2$ for all $\mathbf{v} \in \mathcal{H}$ and for some constants $0 < c_1 \leq c_2 < \infty$. If $c_1 = c_2$, then $\mathbf{F}$ is a tight frame [21, 22].

We use the following definitions for a matrix $\mathbf{X}$ with unit-norm columns: $\delta_s(\mathbf{X})$

denotes the *restricted isometry property* (RIP) constant of order s for $\mathbf{X}$ [23]. We define the *worst-case coherence* of $\mathbf{X}$ as $\mu_1(\mathbf{X}) = \max_{\substack{i,j \\ i \neq j}} \left|\mathbf{x}_i^\top \mathbf{x}_j\right|$. We also define the *order-s cumulative coherence* of $\mathbf{X}$ as

$$\mu_s(\mathbf{X}) \triangleq \max_{|\mathcal{J}| \leq s} \max_{j \notin \mathcal{J}} \|\mathbf{X}_{\mathcal{J}}^\top \mathbf{x}_j\|_1. \tag{1.3}$$

Note that for $s = 1$, the cumulative coherence is equivalent to the worst-case coherence and $\mu_s(\mathbf{X}) \leq s\mu_1(\mathbf{X})$ [24]. For $\mathbf{X} = \bigotimes_{k \in [K]} \mathbf{X}_k$, where $\mathbf{X}_k$'s have unit-norm columns, $\mu_1(\mathbf{X}) = \max_{k \in [K]} \mu_1(\mathbf{X}_k)$ [25, Corollary 3.6] and it can be shown that [12]:

$$\mu_s(\mathbf{X}) \leq \max_{k \in [K]} \mu_{s_k}(\mathbf{X}_k) \Bigg(\prod_{\substack{i \in [K], \\ i \neq k}} (1 + \mu_{s_i - 1}(\mathbf{X}_i)) \Bigg). \tag{1.4}$$

We define the outer product of two vectors of the same dimension, $\mathbf{u}$ and $\mathbf{v}$, as $\mathbf{u} \odot \mathbf{v} = \mathbf{u}\mathbf{v}^\top$ and the inner product between matrices of the same size, $\mathbf{X}$ and $\mathbf{Y}$, as $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{Tr}(\mathbf{X}^\top \mathbf{Y})$. We write $\mathbf{X} \otimes \mathbf{Y}$ for the *Kronecker product* of two matrices $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\mathbf{Y} \in \mathbb{R}^{p \times q}$, defined as

$$\mathbf{X} \otimes \mathbf{Y} = \begin{bmatrix} x_{11}\mathbf{Y} & x_{12}\mathbf{Y} & \dots & x_{1n}\mathbf{Y} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}\mathbf{Y} & x_{m2}\mathbf{Y} & \dots & x_{mn}\mathbf{Y} \end{bmatrix}, \tag{1.5}$$

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]. Given matrices $\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}_1$, and $\mathbf{Y}_2$, where products $\mathbf{X}_1\mathbf{Y}_1$ and $\mathbf{X}_2\mathbf{Y}_2$ can be formed, we have [27]

$$(\mathbf{X}_1 \otimes \mathbf{X}_2)(\mathbf{Y}_1 \otimes \mathbf{Y}_2) = (\mathbf{X}_1\mathbf{Y}_1) \otimes (\mathbf{X}_2\mathbf{Y}_2). \tag{1.6}$$

Given $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\mathbf{Y} \in \mathbb{R}^{p \times n}$, we write $\mathbf{X} * \mathbf{Y}$ for their $mp \times n$ *Khatri-Rao product* [27], defined by

$$\mathbf{X} * \mathbf{Y} = \begin{bmatrix} \mathbf{x}_1 \otimes \mathbf{y}_1 & \mathbf{x}_2 \otimes \mathbf{y}_2 & \dots & \mathbf{x}_n \otimes \mathbf{y}_n \end{bmatrix}. \tag{1.7}$$

This is essentially the column-wise Kronecker product of matrices $\mathbf{X}$ and $\mathbf{Y}$. We also use $\bigotimes_{k\in K}\mathbf{X}_k = \mathbf{X}_1 \otimes \cdots \otimes \mathbf{X}_K$ and $\divideontimes_{k\in K}\mathbf{X}_k = \mathbf{X}_1 * \cdots * \mathbf{X}_K$. For $\mathbf{X} = \bigotimes_{k\in[K]}\mathbf{X}_k$, where $\mathbf{X}_k$'s have unit-norm columns, $\mu_1(\mathbf{X}) = \max_{k\in[K]}\mu_1(\mathbf{X}_k)$ [25, Corollary 3.6].

The separation rank of a matrix $\mathbf{A} \in \mathbb{R}^{\prod_k m_k \times \prod_k p_k}$ is the minimum number R of Kth-order KS matrices $\mathbf{A}^r = \bigotimes_{k=1}^K \mathbf{A}_k^r$ such that $\mathbf{A} = \sum_{r=1}^R \mathbf{A}^r$, where $\mathbf{A}_k^r \in \mathbb{R}^{m_k \times p_k}$.

1.3.1 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 kth-order. The k-rank of a tensor $\underline{\mathbf{X}}$ is defined by $\text{rank}(\mathbf{X}_{(k)})$; trivially, $\text{rank}(\mathbf{X}_{(k)}) \leq p_k$.

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 [5]:

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_k \mathbf{A} \Leftrightarrow \mathbf{Y}_{(k)} = \mathbf{A}\mathbf{X}_{(k)}. \tag{1.8}$$

Tucker Decomposition: The Tucker decomposition decomposes a tensor into a *core tensor* multiplied by a matrix along each mode [5, 19]. We take advantage of the Tucker model since we can relate the Tucker decomposition to the Kronecker representation of tensors [28]. For a tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$ of order K, if $\text{rank}(\mathbf{Y}_{(k)}) \leq p_k$ holds for all $k \in [K]$ then, according to the Tucker model, $\underline{\mathbf{Y}}$ can be decomposed into:

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times_3 \cdots \times_K \mathbf{D}_K, \tag{1.9}$$

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 (1.9) [5]:

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

Since the Kronecker product satisfies $\operatorname{vec}(\mathbf{B}\mathbf{X}\mathbf{A}^\top) = (\mathbf{A} \otimes \mathbf{B}) \operatorname{vec}(\mathbf{X})$, (1.9) is equivalent to

$$\operatorname{vec}(\underline{\mathbf{Y}}) = \big(\mathbf{D}_K \otimes \mathbf{D}_{K-1} \otimes \cdots \otimes \mathbf{D}_1\big) \operatorname{vec}(\underline{\mathbf{X}}), \tag{1.10}$$

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

1.4 Dissertation Outline

The rest of this dissertation is organized as follows. In Chapter 2 of the dissertation, we review some of the theoretical results for DL from vector-valued data and provide some background on DL from tensor data. In Chapter 3 of the dissertation, we focus on the fundamental limits on the sample complexity of estimating dictionaries from tensor data and use an information-theoretical approach to provide general lower bounds on the minimax risk of KS-DL for tensor data. In Chapter 4 of the dissertation, we use tools from real analysis and concentration of measure to derive sufficient conditions for local recovery of coordinate dictionaries comprising a KS dictionary from tensor observations. In Chapter 5 of the dissertation, we extend the KS-DL model to LSR-DL and provide computational algorithms to learn LSR dictionaries using tools from low-rank tensor recovery approaches. In Chapter 6, we study the sparse channel estimation problem in MIMO-OFDM systems and use tools from probability theory and linear algebra to provide theoretical guarantees for channel recovery. We also provide an efficient tensor formulation for the estimation problem and use compressive sensing techniques for tensors to recover the sparse multidimensional channel. Finally in Chapter 7, we provide a summary of the dissertation and a brief overview of future work.

Chapter 2

Background on Dictionary Learning for Vector- and Tensor-Valued Data

During the last decade, dictionary learning has emerged as one of the most powerful methods for data-driven extraction of features from data. While the initial focus on dictionary learning had been from an algorithmic perspective, recent years have seen an increasing interest in understanding the theoretical underpinnings of dictionary learning. Such results help us understand the fundamental limitations of different dictionary learning algorithms. The first part of this chapter focuses on the theoretical aspects of dictionary learning and summarizes existing results that deal with dictionary learning from vector-valued data. These results are primarily stated in terms of lower and upper bounds on the sample complexity of dictionary learning, defined as the number of samples needed to identify or reconstruct the true dictionary underlying data from noiseless or noisy samples, respectively. The second part of this chapter formulates the problem of dictionary learning from tensor-valued data.

2.1 Introduction

There are two major approaches to data representation. In *model-based approaches*, a *predetermined* basis is used to transform data. Such a basis can be formed using predefined transforms such as the Fourier transform [29], wavelets [30], and curvelets [31]. The *data-driven approach* infers transforms from the data to yield efficient representations. Prior works on data representation show that data-driven techniques generally outperform model-based techniques as the learned transformations are tuned to the input signals [3, 4].

Data-driven representations have successfully been used for signal processing and

machine learning tasks such as data compression, recognition, and classification [3, 7, 32]. From a theoretical standpoint, there are several interesting questions surrounding data-driven representations. Assuming there is an unknown generative model forming a "true" representation of data, these questions include: *i)* What algorithms can be used to learn the representation effectively? *ii)* How many data samples are needed to learn the representation? *iii)* What are the fundamental limits on the number of data samples needed to learn the representation? *iv)* How robust are the solutions addressing these questions to parameters such as noise and outliers? In particular, state-of-the-art data representation algorithms have excellent empirical performance but their nonconvex geometry makes analyzing them challenging.

The goal of the first part of this chapter is to provide a brief overview of some of the aforementioned questions for a class of data-driven representation methods known as *dictionary learning* (DL) for vector-valued data. In the second part of the chapter, we provide a formulation of the DL problem for tensor-valued data that is used in the next chapters of the thesis.

2.1.1 Dictionary Learning: A Data-driven Approach to Sparse Representations

Data-driven methods have a long history in representation learning and can be divided into two classes. The first class includes linear methods, which involve transforming (typically vector-valued) data using linear functions to exploit the latent structure in data [3, 33, 34]. From a geometrical point of view, these methods effectively learn a low-dimensional subspace and projection of data onto that subspace, given some constraints. Examples of classical linear approaches for vector-valued data include principal component analysis (PCA) [3], linear discriminant analysis (LDA) [33], and independent component analysis (ICA) [34].

The second class consists of nonlinear methods. Despite the fact that linear representations have historically been preferred over nonlinear methods because of ease of computational complexity, recent advances in available analytical tools and computational power have resulted in an increased interest in nonlinear representation learning.

These techniques have enhanced performance and interpretability compared to linear techniques. In many nonlinear methods, data is transformed into a higher dimensional space, in which it lies on a low dimensional manifold [4, 35–37]. In the world of nonlinear transformations, nonlinearity can take different forms. In manifold-based methods such as diffusion maps, data is projected onto a nonlinear manifold [35]. In kernel (non-linear) PCA, data is projected onto a subspace in a higher dimensional space [36]. Autoencoders encode data based on the desired task [37–39]. DL uses a union of subspaces as the underlying geometric structure and projects input data onto one of the learned subspaces in the union. This leads to sparse representations of the data, which can be represented in the form of an overdetermined matrix multiplied by a sparse vector [4]. Although nonlinear representation methods result in nonconvex formulations, we can often take advantage of the problem structure to guarantee the existence of a unique solution and hence an optimal representation.

DL is known to have slightly higher computational complexity in comparison to linear methods, but it surpasses their performance in applications such as image denoising and inpainting [4], audio processing [6], compressed sensing [40], and data classification [7, 8]. Compared to other nonlinear representation methods, DL provides better interpretability. Furthermore, DL requires less number of samples and is less costly to train compared to autoencoders [41]. More specifically, given input training signals $\mathbf{y} \in \mathbb{R}^m$, the goal in DL is to construct a basis such that $\mathbf{y} \approx \mathbf{D}\mathbf{x}$. Here, $\mathbf{D} \in \mathbb{R}^{m \times p}$ is denoted as the dictionary that has unit-norm columns and $\mathbf{x} \in \mathbb{R}^p$ is the dictionary coefficient vector that has a few nonzero entries. While the initial focus in DL had been on algorithmic development for various problem setups, works in recent years have also provided fundamental analytical results that help us understand the fundamental limits and performance of DL algorithms for vector-valued [1, 24, 42–47].

There are two paradigms in the DL literature: the dictionary can be assumed to be a complete or an overcomplete basis (effectively, a frame [48]). In both cases, columns of the dictionary span the entire space [43]; in complete dictionaries, the dictionary matrix is square ($m = p$), whereas in overcomplete dictionaries the matrix has more columns than rows ($m < p$). In general, overcomplete representations result in more

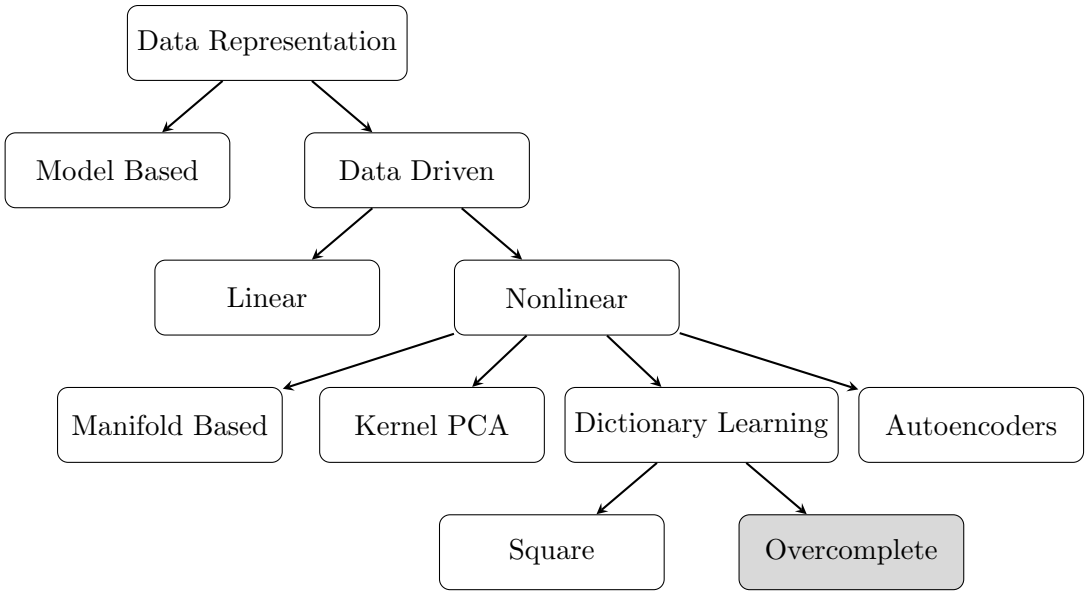


Figure 2.1: A graphical representation of the scope of this chapter in relation to the literature on representation learning.

flexibility to allow both sparse and accurate representations [4].

2.1.2 Chapter Outline

In this chapter, we first summarize key results in learning of overcomplete dictionaries for the case where the data is vector valued. We focus on works that provide fundamental limits on the sample complexity for reliable dictionary estimation, i.e., the number of observations that are necessary to recover the true dictionary that generates the data up to some predefined error. We refer the reader to Fig. 2.1 for a graphical overview of the relationship of this thesis to other themes in representation learning. We focus here only on the problems of *identifiability* and *fundamental limits*; in particular, we do not survey DL algorithms in depth apart from some brief discussion in Sections 2.2 and 2.3. The monograph of Okoudjou [49] discusses algorithms for vector-valued data.

In the second part of the chapter, we provide a formulation of the DL problem for tensor-valued data that is used in the next chapters of the thesis.

2.2 Dictionary Learning for Vector-valued Data

In this section, we address the problem of reliable estimation of dictionaries underlying data that have a single mode, i.e., are vector valued. In particular, we focus on the subject of the sample complexity of the DL problem from two prospectives: *i)* fundamental limits on the sample complexity of DL using *any* DL algorithm, and *ii)* number of samples that are needed for different DL algorithms to reliably estimate a true underlying dictionary that generates the data.

2.2.1 Mathematical Setup

In the conventional vector-valued DL setup, we are given a total number N of vector-valued samples, $\{\mathbf{y}_n \in \mathbb{R}^m\}_{n=1}^N$, that are assumed to be generated from a fixed dictionary, $\mathbf{D}^0$, according to the following model:

$$\mathbf{y}_n = \mathbf{D}^0 \mathbf{x}_n + \mathbf{w}_n, \quad n \in [N]. \tag{2.1}$$

Here, $\mathbf{D}^0 \in \mathbb{R}^{m \times p}$ is a (deterministic) unit-norm frame ($m < p$) that belongs to the following compact set:

$$\mathbf{D}^0 \in \mathcal{D} \triangleq \left\{ \mathbf{D} \in \mathbb{R}^{m \times p}, \|\mathbf{d}_j\|_2 = 1 \ \forall j \in [p] \right\}, \tag{2.2}$$

and is referred to as the *generating*, *true*, or *underlying* dictionary. The vector $\mathbf{x}_n \in \mathbb{R}^p$ is the *coefficient vector* that lies in some set $\mathcal{X} \subseteq \mathbb{R}^p$, and $\mathbf{w}_n \in \mathbb{R}^m$ denotes the random observation noise. Concatenating the observations into a matrix $\mathbf{Y} \in \mathbb{R}^{m \times N}$, their corresponding coefficient vectors into $\mathbf{X} \in \mathbb{R}^{p \times N}$, and noise vectors into $\mathbf{W} \in \mathbb{R}^{m \times N}$, we get the following generative model:

$$\mathbf{Y} = \mathbf{D}^0 \mathbf{X} + \mathbf{W}. \tag{2.3}$$

Various works in the DL literature impose different conditions on the coefficient vectors $\{\mathbf{x}_n\}$ to define the set $\mathcal{X}$. The most common assumption is that $\mathbf{x}_n$ is sparse with one

of several probabilistic models for generating sparse $\mathbf{x}_n$. In contrast to exact sparsity, some works consider approximate sparsity and assume that $\mathbf{x}_n$ satisfies some decay profile [50], while others assume *group sparsity* conditions for $\mathbf{x}_n$ [51]. The latter condition comes up implicitly in DL for tensor data as we discuss in Section 2.3. Similarly, existing works consider a variety of noise models, the most common being Gaussian white noise. Regardless of the assumptions on coefficient and noise vectors, all of these works assume the observations are independent for $n = 1, 2, \ldots, N$.

We are interested here in characterizing when it is possible to recover the true dictionary $\mathbf{D}^0$ from observations $\mathbf{Y}$. There is an inherent ambiguity in dictionary recovery: reordering the columns of $\mathbf{D}^0$ or multiplying any column by -1 yields a dictionary which can generate the same $\mathbf{Y}$ (with appropriately modified $\mathbf{X}$). Thus, each dictionary is equivalent to $2^p p!$ other dictionaries. To measure the distance between dictionaries, we can either define the distance between equivalence classes of dictionaries or consider errors within a local neighborhood of a fixed $\mathbf{D}^0$, where the ambiguity disappears.

The specific criterion that we focus on is sample complexity, defined as the number of observations necessary to recover the true dictionary up to some predefined error. The measure of closeness of the recovered dictionary and the true dictionary can be defined in several ways. One approach is to compare the *representation error* of these dictionaries. Another measure is the mean squared error (MSE) between the estimated and generating dictionary, defined as

$$\mathbb{E}_{\mathbf{Y}} \left\{ d\left(\widehat{\mathbf{D}}(\mathbf{Y}), \mathbf{D}^0\right)^2 \right\}, \tag{2.4}$$

where $d(\cdot, \cdot)$ is some distance metric, and $\widehat{\mathbf{D}}(\mathbf{Y})$ is the recovered dictionary according to observations $\mathbf{Y}$. For example, if we restrict the analysis to a local neighborhood of the generating dictionary, then we can use the Frobenius norm as the distance metric.

We now discuss an optimization approach to solving the dictionary recovery problem. Understanding the objective function within this approach is the key to understanding the sample complexity of DL. Recall that solving the DL problem involves using the observations to estimate a dictionary $\widehat{\mathbf{D}}$ such that $\widehat{\mathbf{D}}$ is close to $\mathbf{D}^0$. In

the ideal case, the objective function involves solving the *statistical risk minimization* problem as follows:

$$\widehat{\mathbf{D}} \in \underset{\mathbf{D}\in\mathcal{D}}{\arg\min}\, \mathbb{E}\left\{ \inf_{\mathbf{x}\in\mathcal{X}} \left\{ \frac{1}{2}\left\|\mathbf{y}-\mathbf{D}\mathbf{x}\right\|_2^2 + \mathcal{R}(\mathbf{x}) \right\} \right\}. \tag{2.5}$$

Here, $\mathcal{R}(\cdot)$ is a regularization operator that enforces the pre-specified structure, such as sparsity, on the coefficient vectors. Typical choices for this parameter include functions of $\|\mathbf{x}\|_0$ or its convex relaxation, $\|\mathbf{x}\|_1$. However, solving (2.5) requires knowledge of exact distributions of the problem parameters as well as high computational power. Hence, works in the literature resort to algorithms that solve the *empirical risk minimization* (ERM) problem [52]:

$$\widehat{\mathbf{D}} \in \underset{\mathbf{D}\in\mathcal{D}}{\arg\min} \left\{ \sum_{n=1}^{N} \inf_{\mathbf{x}_n\in\mathcal{X}} \left\{ \frac{1}{2}\left\|\mathbf{y}_n-\mathbf{D}\mathbf{x}_n\right\|_2^2 + \mathcal{R}(\mathbf{x}_n) \right\} \right\}. \tag{2.6}$$

In particular, to provide analytical results, many estimators solve this problem in lieu of (2.5) and then show that the solution of (2.6) is close to (2.5).

There are a number of computational algorithms that have been proposed to solve (2.6) directly for various regularizers, or indirectly using heuristic approaches. One of the most popular heuristic approaches is the K-SVD algorithm, which can be thought of as solving (2.6) with ℓ_0-norm regularization [4]. There are also other methods such as *method of optimal directions* (MOD) [53] and online DL [8] that solve (2.6) with convex regularizers. While these algorithms have been known to perform well in practice, attention has shifted in recent years to theoretical studies to *i)* find the fundamental limits of solving the statistical risk minimization problem in (2.5), *ii)* determine conditions on objective functions like (2.6) to ensure recovery of the true dictionary, and *iii)* characterize the number of samples needed for recovery using either (2.5) or (2.6). In this chapter, we are also interested in understanding the sample complexity for the DL statistical risk minimization and ERM problems. We summarize such results in the existing literature for the statistical risk minimization of DL in Subsection 2.2.2 and for the ERM problem in Subsection 2.2.3. Because the measure of closeness or error differs between these theoretical results, the corresponding sample complexity bounds

are different.

Remark 2.1. In this section, we assume that the data is available in a batch, centralized setting and the dictionary is deterministic. In the literature, DL algorithms have been proposed for other settings such as streaming data, distributed data, and Bayesian dictionaries [54–57]. Discussions of these scenarios is beyond the scope of this chapter. In addition, some works have looked at ERM problems that are different from (2.6) [7, 8, 58].

2.2.2 Minimax Lower Bounds on the Sample Complexity of DL

In this section, we study the fundamental limits on the accuracy of the dictionary recovery problem that is achievable by *any* DL method in the minimax setting. Specifically, we wish to understand the behavior of the *best estimator* that achieves the lowest *worst-case MSE* among all possible estimators. We define the error of such an estimator as the *minimax risk*, which is formally defined as:

$$\varepsilon^* = \inf_{\widehat{\mathbf{D}}(\mathbf{Y})} \sup_{\mathbf{D}\in\widetilde{\mathcal{D}}} \mathbb{E}_{\mathbf{Y}} \left\{ d\left(\widehat{\mathbf{D}}(\mathbf{Y}), \mathbf{D}\right)^2 \right\}. \tag{2.7}$$

Note that the minimax risk does not depend on any specific DL method and provides a lower bound for the error achieved by any estimator.

The first result we present pertains to lower bounds on the minimax risk, i.e., minimax lower bounds, for the DL problem using the Frobenius norm as the distance metric between dictionaries. The result is based on the following assumption:

A1.1 (Local recovery) The true dictionary lies in a neighborhood of a fixed, known reference dictionary,[1] $\mathbf{D}^* \in \mathcal{D}$, i.e., $\mathbf{D}^0 \in \widetilde{\mathcal{D}}$, where

$$\widetilde{\mathcal{D}} = \left\{ \mathbf{D} | \mathbf{D} \in \mathcal{D}, \|\mathbf{D} - \mathbf{D}^*\|_F \leq r \right\}. \tag{2.8}$$

The range for the neighborhood radius r in (2.8) is $(0, 2\sqrt{p}]$. This conditioning comes

[1] The use of a reference dictionary is an artifact of the proof technique and for sufficiently large r, $\mathcal{D} \approx \widetilde{\mathcal{D}}$.

from the fact that for any $\mathbf{D}, \mathbf{D}' \in \mathcal{D}$, $\|\mathbf{D} - \mathbf{D}'\|_F \leq \|\mathbf{D}\|_F + \|\mathbf{D}'\|_F = 2\sqrt{p}$. By restricting dictionaries to this class, for small enough r, ambiguities that are a consequence of using the Frobenius norm can be prevented. We also point out that any lower bound on ε^* is also a lower bound on the global DL problem.

Theorem 2.1 (Minimax lower bounds [1])**.** *Consider a DL problem for vector-valued data with N i.i.d. observations and true dictionary $\mathbf{D}$ satisfying assumption **A1.1** for some $r \in (0, 2\sqrt{p}]$. Then for any coefficient distribution with mean zero and covariance matrix $\mathbf{\Sigma}_x$, and white Gaussian noise with mean zero and variance σ^2, the minimax risk ε^* is lower bounded as*

$$\varepsilon^* \geq c_1 \min\left\{r^2, \frac{\sigma^2}{N\left\|\mathbf{\Sigma}_x\right\|_2}(c_2 p(m-1)-1)\right\}, \tag{2.9}$$

for some positive constants c_1 and c_2.

Theorem 2.1 holds for both square and overcomplete dictionaries. To obtain this lower bound on the minimax risk, a standard information-theoretic approach is taken in [1] to reduce the dictionary estimation problem to a multiple hypothesis testing problem. In this technique, given fixed $\mathbf{D}^*$ and r, and $L \in \mathbb{N}$, a packing $\mathcal{D}_L = \{\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_L\} \subseteq \widetilde{\mathcal{D}}$ of $\widetilde{\mathcal{D}}$ is constructed. The distance of the packing is chosen to ensure a tight lower bound on the minimax risk. Given observations $\mathbf{Y} = \mathbf{D}_l\mathbf{X} + \mathbf{W}$, where $\mathbf{D}_l \in \mathcal{D}_L$ and index l is chosen uniformly at random from $[L]$, and any estimation algorithm that recovers a dictionary $\widehat{\mathbf{D}}(\mathbf{Y})$, a minimum distance detector can be used to find the recovered dictionary index $\widehat{l} \in [L]$. Then, Fano's inequality can be used to relate the probability of error, i.e., $\mathbb{P}(\widehat{l}(\mathbf{Y}) \neq l)$, to the mutual information between observations and the dictionary (equivalently, the dictionary index l), i.e., $I(\mathbf{Y}; l)$ [59].

Let us assume that r is sufficiently large such that the minimizer of the left hand side of (2.9) is the second term. In this case, Theorem 2.1 states that to achieve any error $\varepsilon \geq \varepsilon^*$, we need the number of samples to be on the order of $N = \Omega\left(\frac{\sigma^2 mp}{\left\|\mathbf{\Sigma}_x\right\|_2 \varepsilon}\right)$. Hence, the lower bound on the minimax risk of DL can be translated to a lower bound on the number of necessary samples, as a function of the desired dictionary error. This can further be interpreted as a lower bound on the sample complexity of the dictionary

recovery problem.

We can also specialize this result to sparse coefficient vectors. Assume $\mathbf{x}_n$ has up to s nonzero elements and the random support of the nonzero elements of $\mathbf{x}_n$ is assumed to be uniformly distributed over the set $\{\mathcal{S} \subseteq [p] : |\mathcal{S}| = s\}$, for $n = [N]$. Assuming that the nonzero entries of $\mathbf{x}_n$ are i.i.d. with variance σ_x^2, we get $\mathbf{\Sigma}_x = (s/p)\sigma_x^2 \mathbf{I}_p$. Therefore, for sufficiently large r, the sample complexity scaling to achieve any error ε becomes $\Omega\left(\frac{\sigma^2 m p^2}{\sigma_x^2 s \varepsilon}\right)$. In this special case, it can be seen that in order to achieve a fixed error ε, the sample complexity scales with the number of degrees of freedom of the dictionary multiplied by number of dictionary columns, i.e., $N = \Omega(mp^2)$. There is also an inverse dependence on sparsity level s. Defining the signal-to-noise-ratio of the observations as $\text{SNR} = \frac{s\sigma_x^2}{m\sigma^2}$, this can be interpreted as an inverse relationship with SNR. Moreover, if all parameters except data dimension, m, are fixed, increasing m requires a linear increase in N. Evidently, this linear relation is limited by the fact that $m \leq p$ has to hold to maintain completeness or overcompleteness of the dictionary: increasing m by a large amount requires increasing p also.

While the tightness of this result remains an open problem, Jung et al. [1] have shown that for a special class of square dictionaries that are perturbations of the identity matrix, and for sparse coefficients following a specific distribution, this result is order-wise tight. In other words, a square dictionary that is perturbed from the identity matrix can be recovered from this sample size order. Although this result does not extend to overcomplete dictionaries, it suggests that the lower bounds may be tight.

Finally, while distance metrics that are invariant to dictionary ambiguities have been used for achievable overcomplete dictionary recovery results [46, 47], obtaining minimax lower bounds for DL using these distance metrics remains an open problem.

In this section, we discussed the number of *necessary* samples for reliable dictionary recovery (sample complexity lower bound). In the next subsection, we focus on achievability results, i.e., the number of *sufficient* samples for reliable dictionary recovery (sample complexity upper bound).

2.2.3 Achievability Results

The preceding lower bounds on minimax risk hold for any estimator or computational algorithm. However, the proofs do not provide an understanding of how to construct effective estimators and provide little intuition about the potential performance of practical estimation techniques. In this section, we direct our attention to explicit reconstruction methods and their sample complexities that ensure reliable recovery of the underlying dictionary. Since these *achievability* results are tied to specific algorithms that are guaranteed to recover the true dictionary, the sample complexity bounds from these results can also be used to derive upper bounds on the minimax risk. As we will see later, there remains a gap between the lower bound and the upper bound on the minimax risk. Alternatively, one can interpret the sample complexity lower bound and upper bound as the number of necessary samples and sufficient samples for reliable dictionary recovery, respectively. In the following, we only focus on *identifiability* results: the estimation procedures are not required to be computationally efficient.

One of the first achievability results for DL were derived in [43, 44] for square matrices. Since then, a number of works have been carried out for overcomplete DL involving vector-valued data [24, 42, 45–47, 50]. These works differ from each other in terms of their assumptions on the true underlying dictionary, the dictionary coefficients, presence or absence of noise and outliers, reconstruction objective function, the distance metric used to measure the accuracy of the solution, and the local or global analysis of the solution. In this section, we summarize a few of these results based on various assumptions on the noise and outliers and provide a brief overview of the landscape of these results in Table 2.1. We begin our discussion with achievability results for DL for the case where $\mathbf{Y}$ is exactly given by $\mathbf{Y} = \mathbf{D}^0\mathbf{X}$, i.e., the noiseless setting.

Before proceeding, we provide an assumption that will be used for the rest of this section. We note that the constants that are used in the presented theorems change from one result to another.

(Random support of sparse coefficient vectors). For any $\mathbf{x}_n$ that has up to s nonzero elements, the support of the nonzero elements of $\mathbf{x}_n$ is assumed to be

distributed uniformly at random over the set $\{\mathcal{S} \subseteq [p] : |\mathcal{S}| = s\}$, for $n = [N]$.

Noiseless Recovery

We begin by discussing the first work that proves local identifiability of the overcomplete DL problem. The objective function that is considered in that work is

$$\left(\widehat{\mathbf{X}}, \widehat{\mathbf{D}}\right) = \underset{\mathbf{D}\in\mathcal{D},\mathbf{X}}{\arg\min} \|\mathbf{X}\|_1 \text{ subject to } \mathbf{Y} = \mathbf{D}\mathbf{X}. \tag{2.10}$$

This result is based on the following set of assumptions:

A2.1 (Gaussian random coefficients). The values of the nonzero entries of $\mathbf{x}_n$'s are independent Gaussian random variables with zero mean and common standard deviation $\sigma_x = \sqrt{p/sN}$.

A2.2 (Sparsity level). The sparsity level satisfies $s \leq \min\left\{c_1/\mu(\mathbf{D}^0), c_2 p\right\}$ for some constants c_1 and c_2.

Theorem 2.2 (Noiseless, local recovery [45])**.** *There exist positive constants c_1, c_2 such that if assumptions **A2.1**–**A2.2** are satisfied for true $(\mathbf{X}, \mathbf{D}^0)$, then $(\mathbf{X}, \mathbf{D}^0)$ is a local minimum of* (2.10) *with high probability.*

The probability in this theorem depends on various problem parameters and implies that $N = \Omega\left(sp^3\right)$ samples are sufficient for the desired solution, i.e., true dictionary and coefficient matrix, to be locally recoverable. The proof of this theorem relies on studying the local properties of (2.10) around its optimal solution and does not require defining a distance metric.

We now present a result that is based on the use of a combinatorial algorithm, which can provably and exactly recover the true dictionary. The proposed algorithm solves the objective function is (2.6) with $\mathcal{R}(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$, where λ is the regularization parameter and the distance metric that is used is the column-wise distance. Specifically, for two dictionaries $\mathbf{D}^1$ and $\mathbf{D}^2$, their column-wise distance is defined as

$$d(\mathbf{d}_j^1, \mathbf{d}_j^2) = \min_{l\in\{-1,1\}} \left\|l\mathbf{d}_j^1 - \mathbf{d}_j^2\right\|_2, j \in [p] \tag{2.11}$$

where $\mathbf{d}_j^1$ and $\mathbf{d}_j^2$ are the jth column of $\mathbf{D}^1$ and $\mathbf{D}^2$, respectively. This distance metric avoids the sign-permutation ambiguity among dictionaries belonging to the same equivalence class. To solve (2.6), Agarwal et al. provide a novel DL algorithm that consists of an initial dictionary estimation stage and an alternating minimization stage to update the dictionary and coefficient vectors [46]. The provided guarantees are based on using this algorithm to update the dictionary and coefficients. The result in Theorem 2.3 is based on the following set of assumptions:

A3.1 (Bounded random coefficients). The nonzero entries of $\mathbf{x}_n$'s are drawn from a zero-mean unit-variance distribution and their magnitude satisfies $x_{\min} \leq |x_{n,i}| \leq x_{\max}$.

A3.2 (Sparsity level). The sparsity level satisfies $s \leq \min\left\{c_1/\sqrt{\mu(\mathbf{D}^0)}, c_2 m^{1/9}, c_3 p^{1/8}\right\}$ for some positive constants c_1, c_2, c_3 that depend on $x_{\min}$, $x_{\max}$, and the spectral norm of $\mathbf{D}^0$.

A3.3 (Dictionary assumptions). The true dictionary has bounded spectral norm, i.e., $\left\|\mathbf{D}^0\right\|_2 \leq c_4\sqrt{p/m}$, for some positive c_4.

Theorem 2.3 (Noiseless, exact recovery [46])**.** *Consider a DL problem with N i.i.d. observations and assume that assumptions **A3.1**–**A3.3** are satisfied. Then, there exists a universal constant c such that for given $\eta > 0$, if*

$$N \geq c\left(\frac{x_{\max}}{x_{\min}}\right)^2 p^2 \log\frac{2p}{\eta}, \tag{2.12}$$

there exists a procedure consisting of an initial dictionary estimation stage and an alternating minimization stage such that after $T = \mathcal{O}(\log(\frac{1}{\varepsilon}))$ iterations of the second stage, with probability at least $1 - 2\eta - 2\eta N^2$, $d(\widehat{\mathbf{d}}_j, \mathbf{d}_j^0) \leq \varepsilon, \forall j \in [p], \forall \varepsilon > 0$.

This theorem guarantees that the true dictionary can be recovered to an arbitrary precision given $N = \Omega(p^2 \log p)$ samples. This result is based on two steps. The first step is guaranteeing an error bound for the initial dictionary estimation step. This step involves using a clustering-style algorithm to approximate the dictionary columns. The second step is proving a local convergence result for the alternating minimization stage.

This step involves improving estimates of the coefficient vectors and the dictionary through Lasso [60] and least-square steps, respectively. More details for this work can be found in the paper by Agarwal et al. [46].

While some other works study the sample complexity of the overcomplete DL problem, they do not take noise into account [45, 46]. Next, we present works that obtain sample complexity results for noisy reconstruction of dictionaries.

Noisy Reconstruction

The next result we discuss is based on the following objective function:

$$\max_{\mathbf{D}\in\mathcal{D}} \frac{1}{N}\sum_{n=1}^{N} \max_{|\mathcal{S}|=s} \|\mathbf{P}_{\mathcal{S}}(\mathbf{D})\mathbf{y}_n\|_2^2\,, \tag{2.13}$$

where $\mathbf{P}_{\mathcal{S}}(\mathbf{D})$ denotes projection of $\mathbf{D}$ onto the span of $\mathbf{D}_{\mathcal{S}} = \{\mathbf{d}_j\}_{j\in\mathcal{S}}$.[2] Here, the distance metric that is used is $d(\mathbf{D}^1,\mathbf{D}^2) = \max_{j\in[p]} \left\|\mathbf{d}_j^1 - \mathbf{d}_j^2\right\|_2$. In addition, the results are based on the following set of assumptions:

A4.1 (Unit-norm tight frame). The true dictionary is a unit-norm tight frame, i.e., for all $\mathbf{v}\in\mathbb{R}^m$ we have $\sum_{j=1}^{p}\left|\langle\mathbf{d}_j^0,\mathbf{v}\rangle\right|^2 = \frac{p\|\mathbf{v}\|_2^2}{m}$.

A4.2 (Lower isometry constant). The lower isometry constant of $\mathbf{D}^0$, defined as $\delta_s(\mathbf{D}^0) \triangleq \max_{|\mathcal{S}|\le s}\delta_{\mathcal{S}}(\mathbf{D}^0)$ with $1-\delta_{\mathcal{S}}(\mathbf{D}^0)$ denoting the minimal eigenvalue of ${\mathbf{D}_{\mathcal{S}}^0}^*\mathbf{D}_{\mathcal{S}}^0$, satisfies $\delta_s(\mathbf{D}^0) \le 1 - \frac{s}{m}$.

A4.3 (Decaying random coefficients). The coefficient vector $\mathbf{x}_n$ is drawn from a symmetric decaying probability distribution ν on the unit sphere S^{p-1}.[3]

---

[2]This objective function can be thought of as a manipulation of (2.6) with the ℓ_0-norm regularizer for the coefficient vectors. See [50, Equation 2] for more details.

[3]A probability measure ν on the unit sphere S^{p-1} is called symmetric if for all measurable sets $\mathcal{X}\subseteq S^{p-1}$, for all sign sequences $\mathbf{l}\in\{-1,1\}^p$ and all permutations $\pi:\{1,\dots,p\}\to\{1,\dots,p\}$, we have

$$\begin{aligned}\nu(\mathbf{l}\mathcal{X}) = \nu(\mathcal{X}),\ \text{where } \mathbf{l}\mathcal{X} = \{(\mathbf{l}_1\mathbf{x}_1,\dots,\mathbf{l}_p\mathbf{x}_p) : \mathbf{x}\in\mathcal{X}\}\,,\text{and}\\ \nu(\pi(\mathcal{X})) = \nu(\mathcal{X}),\ \text{where } \pi(\mathcal{X}) = \left\{\left(\mathbf{x}_{\pi(1)},\dots,\mathbf{x}_{\pi(p)}\right) : \mathbf{x}\in\mathcal{X}\right\}.\end{aligned} \tag{2.14}$$

A4.4 (Bounded random noise). The vector $\mathbf{w}_n$ is a bounded random white noise vector satisfying $\|\mathbf{w}_n\|_2 \leq M_w$ almost surely, $\mathbb{E}\{\mathbf{w}_n\} = \mathbf{0}$ and $\mathbb{E}\{\mathbf{w}_n \mathbf{w}_n{}^*\} = \rho^2 \mathbf{I}_m$.

A4.5 (Maximal projection constraint). Define $\mathbf{c}(\mathbf{x}_n)$ to be the non-increasing rearrangement of the absolute values of $\mathbf{x}_n$. Given a sign sequence $\mathbf{l} \in \{-1, 1\}^p$ and a permutation operator $\pi : [p] \to [p]$, define $\mathbf{c}_{\pi,\mathbf{l}}(\mathbf{x}_n)$ whose ith element is equal to $\mathbf{l}_i \mathbf{c}(\mathbf{x}_n)_{\pi(i)}$ for $i \in [p]$. There exists $\kappa > 0$ such that for $\mathbf{c}(\mathbf{x}_n)$ and $\mathcal{S}_\pi \triangleq \pi^{-1}([s])$, we have

$$\nu\Bigg(\min_{\pi,\mathbf{l}} \Big(\left\|\mathbf{P}_{\mathcal{S}_\pi}(\mathbf{D}^0)\mathbf{D}^0\mathbf{c}_{\pi,\mathbf{l}}(\mathbf{x}_n)\right\|_2 - \max_{|\mathcal{S}|=s, \mathcal{S}\neq\mathcal{S}_\pi} \left\|\mathbf{P}_{\mathcal{S}}(\mathbf{D}^0)\mathbf{D}^0\mathbf{c}_{\pi,\mathbf{l}}(\mathbf{x}_n)\right\|_2 \Big) \geq 2\kappa + 2M_w \Bigg) = 1. \tag{2.15}$$

Theorem 2.4 (Noisy, local recovery [50])**.** *Consider a DL problem with N i.i.d. observations and assume that assumptions **A4.1**–**A4.5** are satisfied. If for some $0 < q < 1/4$, the number of samples satisfies:*

$$2N^{-q} + N^{-2q} \leq \frac{c_1\sqrt{1-\delta_s(\mathbf{D}^0)}}{\sqrt{s}\left(1 + c_2\sqrt{\log\left(\frac{c_3 p\binom{p}{s}}{c_4 s(1-\delta_s(\mathbf{D}^0))}\right)}\right)}, \tag{2.16}$$

then, with high probability, there is a local maximum of (2.13) *within distance at most $2N^{-q}$ of $\mathbf{D}^0$.*

The constants c_1, c_2, c_3 and c_4 in Theorem 2.4 depend on the underlying dictionary, coefficient vectors, and the underlying noise. The proof of this theorem relies on the fact that for the true dictionary and its perturbations, the maximal response, i.e., $\left\|\mathbf{P}_{\mathcal{S}}(\widetilde{\mathbf{D}})\mathbf{D}^0\mathbf{x}_n\right\|_2$,[4] is attained for the set $\mathcal{S} = \mathcal{S}_\pi$ for most signals. A detailed explanation of the theorem and its proof can be found in the paper of Schnass [50].

In order to understand Theorem 2.4, let us set $q \approx \frac{1}{4} - \frac{\log p}{\log N}$. We can then understand this theorem as follows. Given $N/\log N = \Omega(mp^3)$, except with probability $\mathcal{O}(N^{-mp})$, there is a local minimum of (2.13) within distance $\mathcal{O}(pN^{-1/4})$ of the true dictionary.

[4] $\widetilde{\mathbf{D}}$ can be $\mathbf{D}^0$ itself or some perturbation of $\mathbf{D}^0$.

Moreover, since the objective function that is considered in this work is also solved for the K-SVD algorithm, this result gives an understanding of the performance of the K-SVD algorithm. Compared to results with $\mathcal{R}(\mathbf{x})$ being a function of the ℓ_1-norm [45, 46], this result requires the true dictionary to be a tight frame. On the flip side, the coefficient vector in Theorem 2.4 is not necessarily sparse; instead, it only has to satisfy a decaying condition.

Next, we present a result obtained by Arora et al. [47] that is similar to that of Theorem 2.3 in the sense that it uses a combinatorial algorithm that can provably recover the true dictionary given noiseless observations. It further obtains dictionary reconstruction results for the case of noisy observations. The objective function considered in this work is similar to that of the K-SVD algorithm and can be thought of as (2.6) with $\mathcal{R}(\mathbf{x}) = \lambda \|\mathbf{x}\|_0$, where λ is the regularization parameter.

Similar to Agarwal et al. [46], Arora et al. [47] define two dictionaries $\mathbf{D}^1$ and $\mathbf{D}^2$ to be *column-wise ε close* if there exists a permutation π and $l \in \{-1, 1\}$ such that $\left\|\mathbf{d}_j^1 - l\mathbf{d}_{\pi(j)}^2\right\|_2 \leq \varepsilon$. This distance metric captures the distance between equivalent classes of dictionaries and avoids the sign-permutation ambiguity. They propose a DL algorithm that first uses combinatorial techniques to recover the support of coefficient vectors, by clustering observations into overlapping clusters that use the same dictionary columns. To find these large clusters, they provide a clustering algorithm. Then, the dictionary is roughly estimated given the clusters, and the solution is further refined. The provided guarantees are based on using the proposed DL algorithm. In addition, the results are based on the following set of assumptions:

A5.1 (Bounded coefficient distribution). Nonzero entries of $\mathbf{x}_n$ are drawn from a zero-mean distribution and lie in $[-x_{\max}, -1] \cup [1, x_{\max}]$, where $x_{\max} = \mathcal{O}(1)$. Moreover, conditioned on any subset of coordinates in $\mathbf{x}_n$ being nonzero, nonzero values of $x_{n,i}$ are independent from each other. Finally, the distribution has bounded 3-wise moments, i.e., the probability that $\mathbf{x}_n$ is nonzero in any subset $\mathcal{S}$ of 3 coordinates is at most c^3 times $\prod_{i \in \mathcal{S}} \mathbb{P}\{x_{n,i} \neq 0\}$, where $c = \mathcal{O}(1)$.[5]

[5]This condition is trivially satisfied if the set of the locations of nonzero entries of $\mathbf{x}_n$ is a random subset of size s.

A5.2 (Gaussian noise). The $\mathbf{w}_n$'s are independent and follow a spherical Gaussian distribution with standard deviation $\sigma = o(\sqrt{m})$.

A5.3 (Dictionary coherence). The true dictionary is $\widetilde{\mu}$-incoherent, that is, for all $i \neq j$, $\langle \mathbf{d}_i^0, \mathbf{d}_j^0 \rangle \leq \widetilde{\mu}(\mathbf{D}^0)/\sqrt{m}$ and $\widetilde{\mu}(\mathbf{D}^0) = \mathcal{O}(\log(m))$.

A5.4 (Sparsity level). The sparsity level satisfies $s \leq c_1 \min\left\{p^{2/5}, \frac{\sqrt{m}}{\widetilde{\mu}(\mathbf{D}^0)\log m}\right\}$, for some positive constant c_1.

Theorem 2.5 (Noisy, exact recovery [47]). *Consider a DL problem with N i.i.d. observations and assume that assumptions **A5.1**–**A5.4** are satisfied. Provided that*

$$N = \Omega\left(\sigma^2 \varepsilon^{-2} p \log p \left(\frac{p}{s^2} + s^2 + \log\frac{1}{\varepsilon}\right)\right), \tag{2.17}$$

there is a universal constant c_1 and a polynomial-time algorithm that learns the underlying dictionary. With high probability, this algorithm returns $\widehat{\mathbf{D}}$ that is column-wise ε close to $\mathbf{D}^0$.

For desired error ε, the run time of the algorithm and the sample complexity depend on $\log\frac{1}{\varepsilon}$. With the addition of noise, there is also a dependency on ε^{-2} for N, which is inevitable for noisy reconstruction of the true dictionary [47, 50]. In the noiseless setting, this result translates into $N = \Omega\left(p \log p \left(\frac{p}{s^2} + s^2 + \log\frac{1}{\varepsilon}\right)\right)$.

Noisy Reconstruction with Outliers

In some scenarios, in addition to observations $\mathbf{Y}$ drawn from $\mathbf{D}^0$, we encounter observations $\mathbf{Y}_{out}$ that are not generated according to $\mathbf{D}^0$. We call such observations outliers (as opposed to inliers $\mathbf{Y}$). In this case, the observation matrix is $\mathbf{Y}_{obs} = [\mathbf{Y}, \mathbf{Y}_{out}]$, where $\mathbf{Y}$ is the inlier matrix and $\mathbf{Y}_{out}$ is the outlier matrix. In this part, we study the robustness of dictionary identification in the presence of noise and outliers. The following result studies (2.6) with $\mathcal{R}(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$, where λ is the regularization parameter. Here, the Frobenius norm is considered as the distance metric. In addition, the result is based on the following set of assumptions:

A6.1 (Cumulative coherence). The cumulative coherence of the true dictionary $\mathbf{D}^0$ satisfies $\mu_s(\mathbf{D}^0) \leq 1/4$.

A6.2 (Bounded random coefficients). Assume nonzero entries of $\mathbf{x}_n$ are drawn i.i.d. from a distribution with absolute mean $\mathbb{E}\{|x|\}$ and variance $\mathbb{E}\{x^2\}$. We denote $\mathbf{l}_n = \text{sign}(\mathbf{x}_n)$. Dropping the index of $\mathbf{x}_n$ and $\mathbf{l}_n$ for simplicity of notations, the following assumptions are satisfied for the coefficient vector: $\mathbb{E}\{\mathbf{x}_\mathcal{S}\mathbf{x}_\mathcal{S}^T|\mathcal{S}\} = \mathbb{E}\{x^2\}\mathbf{I}_s$, $\mathbb{E}\{\mathbf{x}_\mathcal{S}\mathbf{l}_\mathcal{S}^T|\mathcal{S}\} = \mathbb{E}\{|x|\}\mathbf{I}_s$, $\mathbb{E}\{\mathbf{l}_\mathcal{S}\mathbf{l}_\mathcal{S}^T|\mathcal{S}\} = \mathbf{I}_s$, $\|\mathbf{x}\|_2 \leq M_x$, and $\min_{i\in\mathcal{S}}|x_i| \geq x_{\min}$. We define $\kappa_x \triangleq \frac{\mathbb{E}\{|x|\}}{\sqrt{\mathbb{E}\{x^2\}}}$ as a measure of the flatness of $\mathbf{x}$. Moreover,the following inequality is satisfied:

$$\frac{\mathbb{E}\{x^2\}}{M_x\mathbb{E}\{|x|\}} > \frac{cs}{(1-2\mu_s(\mathbf{D}^0))p}\left(\left\|\mathbf{D}^0\right\|_2+1\right)\left\|\mathbf{D}^{0T}\mathbf{D}^0-\mathbf{I}\right\|_F, \tag{2.18}$$

where c is a positive constant.

A6.3 (Regularization parameter). The Regularization parameter satisfies $\lambda \leq x_{\min}/4$.

A6.4 (Bounded random noise). Assume nonzero entries of $\mathbf{w}_n$ are drawn i.i.d. from a distribution with mean 0 and variance $\mathbb{E}\{w^2\}$. Dropping the index of vectors for simplicity, $\mathbf{w}$ is a bounded random white noise vector satisfying $\mathbb{E}\{\mathbf{w}\mathbf{w}^T|\mathcal{S}\} = \mathbb{E}\{w^2\}\mathbf{I}_m$, $\mathbb{E}\{\mathbf{w}\mathbf{x}^T|\mathcal{S}\} = \mathbb{E}\{\mathbf{w}\mathbf{l}^T|\mathcal{S}\} = \mathbf{0}$, and $\|\mathbf{w}\|_2 \leq M_w$. Furthermore, denoting $\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\{|x|\}}$:

$$\frac{M_w}{M_x} \leq \frac{7}{2}\left(c_{\max}-c_{\min}\right)\bar{\lambda}, \tag{2.19}$$

where $c_{\min}$ and $c_{\max}$ depend on problem parameters such as s, coefficient distribution, and $\mathbf{D}^0$.

A6.5 (Sparsity level) The sparsity level satisfies $s \leq \frac{p}{16\left(\|\mathbf{D}^0\|_2+1\right)^2}$.

A6.6 (Radius range) The error radius $\varepsilon > 0$ satisfies $\varepsilon \in \left(\bar{\lambda}c_{\min}, \bar{\lambda}c_{\max}\right)$.

A6.7 (Outlier energy). Given inlier matrix $\mathbf{Y} = \{\mathbf{y}_n\}_{n=1}^N$ and outlier matrix $\mathbf{Y}_{out} =$

$\{\mathbf{y'}_n\}_{n=1}^{N_{out}}$, the energy of $\mathbf{Y}_{out}$ satisfies

$$\frac{\|\mathbf{Y}_{out}\|_{1,2}}{N} \leq \frac{c_1 \varepsilon \sqrt{s} \mathbb{E}\left\{\|\mathbf{x}\|_2^2\right\}}{\bar{\lambda} \mathbb{E}\{|x|\}} \left(\frac{A^0}{p}\right)^{3/2} \left[\frac{1}{p}\left(1 - \frac{c_{\min}\bar{\lambda}}{\varepsilon}\right) - c_2\sqrt{\frac{mp+\eta}{N}}\right], \tag{2.20}$$

where $\|\mathbf{Y}_{out}\|_{1,2}$ denotes the sum of the ℓ_2-norms of the columns of $\mathbf{Y}_{out}$, c_1 and c_2 are positive constants, independent of parameters, and A^0 is the lower frame bound of $\mathbf{D}^0$, i.e., $A^0 \|\mathbf{v}\|_2^2 \leq \left\|\mathbf{D}^{0\top}\mathbf{v}\right\|_2^2$ for any $\mathbf{v} \in \mathbb{R}^m$.

Theorem 2.6 (Noisy with outliers, local recovery [24])**.** *Consider a DL problem with N i.i.d. observations and assume that assumptions **A6.1**–**A6.6** are satisfied. Suppose*

$$N > c_0(mp+\eta)p^2 \left(\frac{M_x^2}{\mathbb{E}\left\{\|\mathbf{x}\|_2^2\right\}}\right)^2 \left(\frac{\varepsilon + \left(\frac{M_w}{M_x} + \bar{\lambda}\right) + \left(\frac{M_w}{M_x} + \bar{\lambda}\right)^2}{\varepsilon - c_{\min}\bar{\lambda}}\right), \tag{2.21}$$

*then with probability at least $1 - 2^{-\eta}$, (2.6) admits a local minimum within distance ε of $\mathbf{D}^0$. In addition, this result is robust to the addition of outlier matrix $\mathbf{Y}_{out}$, provided that the assumption in **A6.7** is satisfied.*

The proof of this theorem relies on using the Lipschitz continuity property of the objective function in (2.6) with respect to the dictionary and sample complexity analysis using Rademacher averages and Slepian's Lemma [61]. Theorem 2.6 implies that

$$N = \Omega\left(\left(mp^3 + \eta p^2\right)\left(\frac{M_w}{M_x \varepsilon}\right)^2\right) \tag{2.22}$$

samples are sufficient for the existence of a local minimum within distance ε of true dictionary $\mathbf{D}^0$, with high probability. In the noiseless setting, this result translates into $N = \Omega\left(mp^3\right)$, and sample complexity becomes independent of the radius ε. Furthermore, this result applies to overcomplete dictionaries with dimensions $p = \mathcal{O}(m^2)$.

2.2.4 Summary of Results

In this section, we have discussed DL minimax risk lower bounds [1] and achievability results [24, 45–47, 50]. These results differ in terms of the distance metric they use. A summary of the general scaling of the discussed results for sample complexity of (overcomplete) DL are provided in Table 2.1. We note that these are general scalings that ignore other technicalities. Here, the provided sample complexity results depend on the present or absence of noise and outliers. All the presented results require the underlying dictionary satisfies incoherence conditions in some way. For a one-to-one comparison of these results, the bounds for the case of absence of noise and outliers can be compared. A detailed comparison of the noiseless recovery for square and overcomplete dictionaries can be found in [24, Table I]. While dictionary identifiability has been well studied for vector-valued data, there remains a gap between the upper and lower bounds on the sample complexity. The lower bound presented in Theorem 2.1 is for the case of a particular distance metric, i.e., the Frobenius norm, whereas the presented achievability results in Theorems 2.2–2.6 are based on a variety of distance metrics. Restricting the distance metric to the Frobenius norm, we still observe a gap of order p between the sample complexity lower bound in Theorem 2.1 and upper bound in Theorem 2.6. The partial converse result for square dictionaries that is provided in [1] shows that the lower bound is achievable for square dictionaries close to the identity matrix. For more general square matrices, however, the gap may be significant: either improved algorithms can achieve the lower bounds or the lower bounds may be further tightened. For overcomplete dictionaries the question of whether the upper bound or lower bound is tight remains open. For metrics other than the Frobenius norm, the bounds are incomparable, making it challenging to assess the tightness of many achievability results.

Finally, the works reported in Table 2.1 differ significantly in terms of the mathematical tools they use. Each approach yields a different insight into the structure of the DL problem. However, there is no unified analytical framework encompassing all of these perspectives.

Table 2.1: Summary of the sample complexity results for overcomplete DL of various works

Reference	Jung et al. [1]	Geng et al. [45]	Agarwal et al. [46]	Schnass et al. [50]	Arora et al. [47]	Gribonval et al. [24]
Distance Metric	$\left\|\mathbf{D}^1-\mathbf{D}^2\right\|_F$	–	$\min_{l\in\{\pm1\}}\left\|\mathbf{d}_j^1-l\mathbf{d}_j^2\right\|_2$	$\max_j\left\|\mathbf{d}_j^1-\mathbf{d}_j^2\right\|_2$	$\min_{l\in\{\pm1\},\pi}\left\|\mathbf{d}_j^1-l\mathbf{d}_{\pi(j)}^2\right\|_2$	$\left\|\mathbf{D}^1-\mathbf{D}^2\right\|_F$
Regularizer	ℓ_0	ℓ_1	ℓ_1	ℓ_1	ℓ_0	ℓ_1
Sparse Coefficient Distribution	nonzero i.i.d zero-mean, variance σ_x^2	nonzero i.i.d. $\sim\mathcal{N}(0,\sigma_x)$	nonzero zero-mean unit-variance $x_{\min}\leq\|x_i\|\leq x_{\max}$	symmetric decaying (non-sparse)	nonzero zero-mean $x_i\in\pm[1,x_{\max}]$	nonzero $\|x_i\|>x_{\min}$, $\|\mathbf{x}\|_2\leq M_x$
Sparsity Level	–	$\mathcal{O}(\min\{1/\mu,p\})$	$\mathcal{O}(\min\{1/\sqrt{\mu},m^{1/9},p^{1/8}\})$	$\mathcal{O}(1/\mu)$	$\mathcal{O}(\min\{1/(\mu\log m),p^{2/5}\})$	$\mathcal{O}(m)$
Noise Distribution	i.i.d.$\sim\mathcal{N}(0,\sigma)$	–	–	$\mathbb{E}\{\mathbf{w}\}=\mathbf{0}$ $\mathbb{E}\{\mathbf{w}\mathbf{w}^*\}=\rho^2\mathbf{I}_m$ $\|\mathbf{w}\|_2\leq M_w$	i.i.d.$\sim\mathcal{N}(0,\sigma)$	$\mathbb{E}\{\mathbf{w}\mathbf{w}^\top\|\mathcal{S}\}=\mathbb{E}\{w^2\}\mathbf{I}_m$ $\mathbb{E}\{\mathbf{w}\mathbf{x}^\top\|\mathcal{S}\}=\mathbf{0}$, $\|\mathbf{w}\|_2\leq M_w$
Outlier	–	–	–	–	–	Robust
Local-Global	Local	Local	Global	Local	Global	Local
Sample Complexity	$\frac{mp^2}{\varepsilon^2}$	sp^3	$p^2\log p$	mp^3	$\frac{p}{\varepsilon^2}\log p(p/s^2+s^2+\log\frac{1}{\varepsilon})$	$\frac{mp^3}{\varepsilon^2}$

2.3 Dictionary Learning for Tensors

Many of today's data are collected using various sensors and tend to have a multidimensional/tensor structure. To find representations of tensor data using DL, one can follow two paths. A naive approach is to vectorize tensor data and use traditional vectorized representation learning techniques. A better approach is to take advantage of the multidimensional structure of data to learn representations that are specific to tensor data. While the main focus of the literature on representation learning has been on the former approach, recent works have shifted focus to the latter approaches [11, 14, 16, 62]. These works use various tensor decompositions to decompose tensor data into smaller components. The representation learning problem can then be reduced to learning the components that represent different modes of the tensor. This results in reduction in the number of degrees of freedom in the learning problem, due to the fact that the dimensions of the representations learned for each mode are significantly smaller than the dimensions of the representation learned for the vectorized tensor. Consequently, this approach gives rise to compact and efficient representation of tensors.

To understand the fundamental limits of DL for tensor data, one can use the sample complexity results in Section 2.2, which are a function of the underlying dictionary dimensions. However, considering the reduced number of degrees of freedom in the tensor DL problem compared to vectorized DL, this problem should be solvable with a smaller number of samples. In the next sections of this thesis, we formalize this intuition and address the problem of reliable estimation of dictionaries underlying tensor data. Similar to the previous section, we will focus on the subject of sample complexity of the DL problem from two prospectives; *i)* fundamental limits on the sample complexity of DL for tensor data using any DL algorithm, and *ii)* number of samples that are needed for different DL algorithms to reliably estimate the true dictionary from which the tensor data is generated.

2.3.1 Mathematical Setup

In this chapter, we consider the Tucker decomposition due to the following reasons: *i)* it represents a sequence of independent transformations, i.e., factor matrices, for different data modes, and *ii)* Kronecker-structured matrices have successfully been used for data representation in applications such as magnetic resonance imaging, hyperspectral imaging, video acquisition, and distributed sensing [11, 16].

2.3.2 Kronecker-structured Dictionary Learning (KS-DL)

In order to state the main results of this section, we begin with a generative model for tensor data based on Tucker decomposition. Specifically, we assume we have access to a total number of N tensor observations, $\underline{\mathbf{Y}}_n \in \mathbb{R}^{m_1 \times \cdots \times m_K}$, that are generated according to the following model:[6]

$$\operatorname{vec}(\underline{\mathbf{Y}}_n) = \left(\mathbf{D}_1^0 \otimes \mathbf{D}_2^0 \otimes \cdots \otimes \mathbf{D}_K^0\right) \operatorname{vec}(\underline{\mathbf{X}}_n) + \operatorname{vec}(\underline{\mathbf{W}}_n), \quad n = 1, \ldots, N. \tag{2.23}$$

Here, $\{\mathbf{D}_k^0 \in \mathbb{R}^{m_k \times p_k}\}_{k=1}^K$ are the true fixed *coordinate dictionaries*, $\underline{\mathbf{X}}_n \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the coefficient tensor, and $\underline{\mathbf{W}}_n \in \mathbb{R}^{m_1 \times \cdots \times m_K}$ is the underlying noise tensor. In this case, the true dictionary $\mathbf{D}^0 \in \mathbb{R}^{m \times p}$ is Kronecker-structured (KS) and has the form

$$\begin{aligned} \mathbf{D}^0 = \bigotimes_k \mathbf{D}_k^0, \quad m = \prod_{k=1}^K m_k \quad \text{and} \quad p = \prod_{k=1}^K p_k, \\ \text{where} \quad \mathbf{D}_k^0 \in \mathcal{D}_k = \left\{\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}, \|\mathbf{d}_{k,j}\|_2 = 1 \ \forall j \in [p_k]\right\}. \end{aligned} \tag{2.24}$$

We define the set of KS dictionaries as

$$\mathcal{D}_{KS} = \left\{\mathbf{D} \in \mathbb{R}^{m \times p} : \mathbf{D} = \bigotimes_k \mathbf{D}_k, \mathbf{D}_k \in \mathcal{D}_k \ \forall k \in [K]\right\}. \tag{2.25}$$

Comparing (2.23) to the traditional formulation in (2.1), it can be seen that KS-DL also involves vectorizing the observation tensor, but it has the main difference that the

[6] We have reindexed $\mathbf{D}_k$'s here for simplicity of notation.

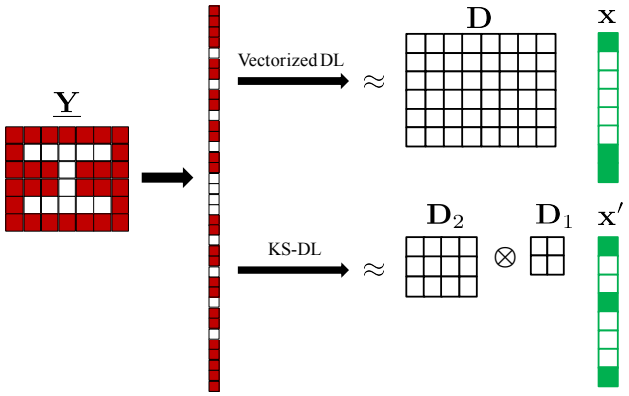


Figure 2.2: Illustration of the distinctions of KS-DL versus vectorized DL for a 2nd-order tensor: both vectorize the observation tensor, but the structure of the tensor is exploited in the KS dictionary, leading to the learning of two coordinate dictionaries with reduced number of parameters compared to the dictionary learned in vectorized DL.

structure of the tensor is captured in the underlying KS dictionary. An illustration of this for a 2nd-order tensor is shown in Figure 2.2. Similar to (2.3), we can stack the vectorized observations, $\mathbf{y}_n = \text{vec}(\underline{\mathbf{Y}}_n)$, vectorized coefficient tensors, $\mathbf{x}_n = \text{vec}(\underline{\mathbf{X}}_n)$, and vectorized noise tensors, $\mathbf{w}_n = \text{vec}(\underline{\mathbf{W}}_n)$, in columns of $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{W}$, respectively. We now discuss the role of sparsity in coefficient tensors for dictionary learning. While in vectorized DL it is usually assumed that the random support of nonzero entries of $\mathbf{x}_n$ is uniformly distributed, there are two different definitions of the random support of $\underline{\mathbf{X}}_n$ for tensor data:

1) Random sparsity: The random support of $\mathbf{x}_n$ is uniformly distributed over the set $\{\mathcal{S} \subseteq [p] : |\mathcal{S}| = s\}$.

2) Separable sparsity: The random support of $\mathbf{x}_n$ is uniformly distributed over the set $\mathcal{S}$ that is related to $\{\mathcal{S}_1 \times \ldots \mathcal{S}_K : \mathcal{S}_k \subseteq [p_k], |\mathcal{S}_k| = s_k\}$ via lexicographic indexing. Here, $s = \prod_k s_k$.

Separable sparsity requires nonzero entries of the coefficient tensor to be grouped in blocks. This model also implies that the columns of $\mathbf{Y}_{(k)}$ have s_k-sparse representations with respect to $\mathbf{D}_k^0$ [28].

The aim in KS-DL is to estimate coordinate dictionaries, $\widehat{\mathbf{D}}_k$'s, such that they are

close to $\mathbf{D}_k^0$'s. In this scenario, the statistical risk minimization problem has the form:

$$\left(\widehat{\mathbf{D}}_1,\ldots,\widehat{\mathbf{D}}_K\right) \in \underset{\{\mathbf{D}_k\in\mathcal{D}_k\}_{k=1}^K}{\arg\min}\ \mathbb{E}\left\{\inf_{\mathbf{x}\in\mathcal{X}}\left\{\frac{1}{2}\left\|\mathbf{y}-\left(\bigotimes_k \mathbf{D}_k\right)\mathbf{x}\right\|_2^2+\mathcal{R}(\mathbf{x})\right\}\right\}, \tag{2.26}$$

and the ERM problem is formulated as:

$$\left(\widehat{\mathbf{D}}_1,\ldots,\widehat{\mathbf{D}}_K\right) \in \underset{\{\mathbf{D}_k\in\mathcal{D}_k\}_{k=1}^K}{\arg\min}\ \left\{\sum_{n=1}^N \inf_{\mathbf{x}_n\in\mathcal{X}}\left\{\frac{1}{2}\left\|\mathbf{y}_n-\left(\bigotimes_k \mathbf{D}_k\right)\mathbf{x}_n\right\|_2^2+\mathcal{R}(\mathbf{x}_n)\right\}\right\}, \tag{2.27}$$

where $\mathcal{R}(.)$ is a regularization operator on the coefficient vectors. Various KS-DL algorithms have been proposed that solve (2.27) heuristically by means of optimization tools such as alternative minimization [16] and tensor rank minimization [63], and by taking advantage of techniques in tensor algebra such as the higher-order SVD for tensors [64].

In the case of theory for KS-DL, the notion of closeness can have two interpretations. One is the distance between the true KS dictionary and the recovered KS dictionary, i.e., $d\left(\widehat{\mathbf{D}}(\mathbf{Y}),\mathbf{D}^0\right)$. The other is the distance between each true coordinate dictionary and the corresponding recovered coordinate dictionary, i.e., $d\left(\widehat{\mathbf{D}}_k(\mathbf{Y}),\mathbf{D}_k^0\right)$. While small recovery errors for coordinate dictionaries imply a small recovery error for the KS dictionary, the other side of the statement does not necessarily hold. Hence, the latter notion is of importance when we are interested in recovering the structure of the KS dictionary.

In Chapters 3 and 4, we focus on the sample complexity of the KS-DL problem. The questions that we address in these chapters are *i)* What are the fundamental limits of solving the statistical risk minimization problem in (2.26)? *ii)* Under what kind of conditions do objective functions like (2.27) recover the true coordinate dictionaries and how many samples do they need for this purpose? *iii)* How do these limits compare to their vectorized DL counterparts? Addressing these question will help in understanding the benefits of KS-DL for tensor data.

Chapter 3

Fundamental Limits on the Minimax Risk of Kronecker-structured Dictionary Learning

In this chapter, we study the fundamental limits on the sample complexity of estimating dictionaries for tensor data. The specific focus of this chapter is on Kth-order tensor data and the case where the underlying dictionary can be expressed in terms of K smaller dictionaries. It is assumed the data are generated by linear combinations of these structured dictionary atoms and observed through white Gaussian noise. This chapter first provides a general lower bound on the minimax risk of dictionary learning for such tensor data and then adapts the proof techniques for specialized results in the case of sparse and sparse-Gaussian linear combinations. A partial converse is provided for the case of 2nd-order tensor data to show that the bounds in this chapter can be tight. This involves developing an algorithm for learning highly-structured dictionaries from noisy tensor data. Finally, numerical experiments highlight the advantages associated with explicitly accounting for tensor data structure during dictionary learning.[1]

3.1 Introduction

In traditional DL literature, multidimensional data are converted into one-dimensional data by vectorizing the signals. Such approaches can result in poor sparse representations because they neglect the multidimensional structure of the data [68]. This suggests that it might be useful to keep the original tensor structure of multidimensional data for efficient DL and reliable subsequent processing.

[1]The results presented in this chapter have been published in Proceedings of 2016 IEEE International Symposium on Information Theory [65], Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing [66], and IEEE Transactions on Information Theory [67].

There have been several algorithms proposed in the literature that can be used to learn structured dictionaries for multidimensional data [14–16, 62, 68–73]. In [14], a Riemannian conjugate gradient method combined with a nonmonotone line search is used to learn structured dictionaries. Other structured DL works rely on various tensor decomposition methods such as the Tucker decomposition [15, 16, 19, 62, 70, 71], the CP decomposition [18, 73], the HOSVD decomposition [64, 69], the t-product tensor factorization [72], and the tensor-SVD [68, 74]. Furthermore learning sums of structured dictionaries can be used to represent tensor data [70].

In this chapter, our focus is on theoretical understanding of the fundamental limits of DL algorithms that explicitly account for the tensor structure of data in terms of *Kronecker structured* (KS) dictionaries. KS matrices have successfully been used for data representation in hyperspectral imaging, video acquisition, and distributed sensing[11].

To the best of our knowledge, none of the prior works on KS DL [14–16, 69, 70] provide an understanding of the sample complexity of KS-DL algorithms. In contrast, we provide lower bounds on the minimax risk of estimating KS dictionaries from tensor data using *any* estimator. These bounds not only provide means of quantifying the performance of existing KS-DL algorithms, but they also hint at the potential benefits of explicitly accounting for tensor structure of data during DL.

3.1.1 Our Contributions

Our first result is a general lower bound for the mean squared error (MSE) of estimating KS-dictionaries consisting of $K \geq 2$ coordinate dictionaries that sparsely represent Kth-order tensor data. Here, we define the minimax risk to be the worst-case MSE that is attainable by the best dictionary estimator. Our approach uses the standard procedure for lower bounding the minimax risk in nonparametric estimation by connecting it to the maximum probability of error on a carefully constructed multiple hypothesis testing problem [59, 75]: the technical challenge is in constructing an appropriate set of hypotheses. In particular, consider a dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ consisting of the Kronecker product of K coordinate dictionaries $\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}, k \in \{1, \ldots, K\}$, where

$m = \prod_{k=1}^{K} m_k$ and $p = \prod_{k=1}^{K} p_k$, that is generated within the radius r neighborhood (taking the Frobenius norm as the distance metric) of a fixed reference dictionary. Our analysis shows that given a sufficiently large r and keeping some other parameters constant, a sample complexity of $N = \Omega(\sum_{k=1}^{K} m_k p_k)$ is necessary for reconstruction of the true dictionary up to a given estimation error. We also provide minimax bounds on the KS-DL problem that hold for the following distributions for the coefficient vectors $\{\mathbf{x}_n\}$:

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

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

3.1.2 Relationship to Previous Work

In terms of relation to prior work, theoretical insights into the problem of DL have either focused on specific algorithms for non-KS dictionaries [24, 42, 46, 47, 50, 58, 76] or lower bounds on minimax risk of DL for one-dimensional data [1, 77]. The former works provide sample complexity results for reliable dictionary estimation based on appropriate minimization criteria. Specifically, given a probabilistic model for sparse coefficients and a finite number of samples, these works find a local minimizer of a nonconvex objective function and show that this minimizer is a dictionary within a given distance of the true dictionary [24, 50, 58]. In contrast, Jung et al. [1, 77] provide minimax lower bounds for DL from one-dimensional data under several coefficient

vector distributions and discuss a regime where the bounds are tight in the scaling sense for some signal-to-noise (SNR) values. In particular, for a given dictionary $\mathbf{D}$ and sufficiently large neighborhood radius r, they show that $N = \Omega(mp)$ samples are required for reliable recovery of the dictionary up to a prescribed MSE within its local neighborhood. However, in the case of tensor data, their approach does not exploit the structure in the data, whereas our goal is to show how structure can potentially yield a lower sample complexity in the DL problem.

To provide lower bounds on the minimax risk of KS DL, we adopt the same general approach that Jung et al. [1, 77] use for the vector case. They use the standard approach of connecting the estimation problem to a multiple-hypothesis testing problem and invoking Fano's inequality [59]. We construct a family of KS dictionaries which induce similar observation distributions but have a minimum separation from each other. By explicitly taking into account the Kronecker structure of the dictionaries, we show that the sample complexity satisfies a lower bound of $\Omega(\sum_{k=1}^{K} m_k p_k)$ compared to the $\Omega(mp)$ bound from vectorizing the data [1]. Although our general approach is similar to that in [1], there are fundamental differences in the construction of the KS dictionary class and analysis of the minimax risk. This generalizes our preliminary work [65] from 2nd-order to Kth-order and provides a comprehensive analysis of the KS dictionary class construction and minimax lower bounds.

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

We next introduce a KS-DL algorithm for 2nd-order tensor data and show that in this case, one of the provided minimax lower bounds is achievable under certain

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

conditions. We also conduct numerical experiments that demonstrate the empirical performance of the algorithm relative to the MSE upper bound and in comparison to the performance of a non-KS DL algorithm [1].

3.2 Problem Formulation

We assume the observations are Kth-order tensors $\underline{\mathbf{Y}}_n \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$. According to the Tucker model, given *coordinate dictionaries* $\mathbf{D}_k^0 \in \mathbb{R}^{m_k \times p_k}$, a *coefficient tensor* $\underline{\mathbf{X}}_n \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_K}$, and a *noise tensor* $\underline{\mathbf{W}}_n$, we can write $\mathbf{y}_n \triangleq \mathrm{vec}(\underline{\mathbf{Y}}_n)$ using (1.10) as[3]

$$\mathbf{y}_n = \left(\bigotimes_{k \in [K]} \mathbf{D}_k^0 \right) \mathbf{x}_n + \mathbf{w}_n, \tag{3.1}$$

where $\mathbf{x}_n \triangleq \mathrm{vec}(\underline{\mathbf{X}}_n)$ and $\mathbf{w}_n \triangleq \mathrm{vec}(\underline{\mathbf{W}}_n)$. Let

$$m = \prod_{k \in [K]} m_k \quad \text{and} \quad p = \prod_{k \in [K]} p_k. \tag{3.2}$$

Concatenating N i.i.d. noisy observations $\{\mathbf{y}_n\}_{n=1}^N$, which are realizations according to the model (3.1), into $\mathbf{Y} \in \mathbb{R}^{m \times N}$, we obtain

$$\mathbf{Y} = \mathbf{D}^0 \mathbf{X} + \mathbf{W}, \tag{3.3}$$

where $\mathbf{D}^0 \triangleq \bigotimes_{k \in [K]} \mathbf{D}_k^0$ is the unknown KS dictionary, $\mathbf{X} \in \mathbb{R}^{p \times N}$ is a coefficient matrix consisting of i.i.d. random coefficient vectors with known distribution that has zero-mean and covariance matrix $\mathbf{\Sigma}_x$, and $\mathbf{W} \in \mathbb{R}^{m \times N}$ is assumed to be additive white Gaussian noise (AWGN) with zero mean and variance σ^2.

Our main goal in this chapter is to derive necessary conditions under which the KS dictionary $\mathbf{D}^0$ can possibly be learned from the noisy observations given in (3.3). We assume the true KS dictionary $\mathbf{D}^0$ consists of unit-norm columns and we carry out local analysis. That is, the true KS dictionary $\mathbf{D}^0$ is assumed to belong to a neighborhood

[3] We have reindexed $\mathbf{D}_k$'s in (1.10) for ease of notation.

around a fixed (normalized) reference KS dictionary

$$\mathbf{D}^* = \bigotimes_{k\in[K]} \mathbf{D}_k^*, \tag{3.4}$$

and $\mathbf{D}^* \in \mathcal{D}_{KS}$, where $\mathcal{D}_{KS}$ is defined in (2.25). We assume the true generating KS dictionary $\mathbf{D}^0$ belongs to a neighborhood around $\mathbf{D}^*$:

$$\mathbf{D}^0 \in \mathcal{X}(\mathbf{D}^*, r) \triangleq \left\{\mathbf{D}' \in \mathcal{D}_{KS} : \left\|\mathbf{D}' - \mathbf{D}^*\right\|_F < r\right\} \tag{3.5}$$

for some fixed radius r.[4] Note that $\mathbf{D}^*$ appears in the analysis as an artifact of our proof technique to construct the dictionary class. In particular, if r is sufficiently large, then $\mathcal{X}(\mathbf{D}^*, r) \approx \mathcal{D}_{KS}$ and effectively $\mathbf{D} \in \mathcal{D}_{KS}$.

3.2.1 Minimax Risk

We are interested in lower bounding the minimax risk for estimating $\mathbf{D}^0$ based on observations $\mathbf{Y}$, which is defined as the worst-case mean squared error (MSE) that can be obtained by the best KS dictionary estimator $\widehat{\mathbf{D}}(\mathbf{Y})$. That is,

$$\varepsilon^* = \inf_{\widehat{\mathbf{D}}} \sup_{\mathbf{D}^0\in\mathcal{X}(\mathbf{D}^*,r)} \mathbb{E}_{\mathbf{Y}} \left\{ \left\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}^0\right\|_F^2 \right\}, \tag{3.6}$$

where $\widehat{\mathbf{D}}(\mathbf{Y})$ can be estimated using any KS-DL algorithm. In order to lower bound this minimax risk ε^*, we employ a standard reduction to the multiple hypothesis testing used in the literature on nonparametric estimation [59, 75]. This approach is equivalent to generating a KS dictionary $\mathbf{D}_l$ uniformly at random from a carefully constructed class $\mathcal{D}_L = \{\mathbf{D}_1, \ldots, \mathbf{D}_L\} \subseteq \mathcal{X}(\mathbf{D}^*, r), L \geq 2$, for a given ($\mathbf{D}^*$, r). To ensure a tight lower bound, we must construct $\mathcal{D}_L$ such that the distance between any two dictionaries in $\mathcal{D}_L$ is large but the hypothesis testing problem is hard; that is, two distinct dictionaries $\mathbf{D}_l$ and $\mathbf{D}_{l'}$ should produce similar observations. Specifically, for $l, l' \in [L]$, and given

[4] Note that our results hold with the unit-norm condition enforced only on $\mathbf{D}^0$ itself, and not on the subdictionaries $\mathbf{D}_k^0$. Nevertheless, we include this condition in the dictionary class for the sake of completeness as it also ensures uniqueness of the subdictionaries (factors of a K-fold Kronecker product can exchange scalars γ_k freely without changing the product as long as $\prod_{k\in[K]} \gamma_k = 1$).

error $\varepsilon \geq \varepsilon^*$, we desire a construction such that

$$\forall l \neq l', \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F \geq 2\sqrt{\gamma\varepsilon} \quad \text{and} \quad D_{KL}\left(f_{\mathbf{D}_l}(\mathbf{Y})||f_{\mathbf{D}_{l'}}(\mathbf{Y})\right) \leq \alpha_L, \tag{3.7}$$

where $D_{KL}\left(f_{\mathbf{D}_l}(\mathbf{Y})||f_{\mathbf{D}_{l'}}(\mathbf{Y})\right)$ denotes the Kullback-Leibler (KL) divergence between the distributions of observations based on $\mathbf{D}_l \in \mathcal{D}_L$ and $\mathbf{D}_{l'} \in \mathcal{D}_L$, while γ, α_L, and ε are non-negative parameters. Observations $\mathbf{Y} = \mathbf{D}_l\mathbf{X} + \mathbf{W}$ in this setting can be interpreted as channel outputs that are used to estimate the input $\mathbf{D}_l$ using an arbitrary KS dictionary algorithm that is assumed to achieve the error ε. Our goal is to detect the correct generating KS dictionary index l. For this purpose, a minimum distance detector is used:

$$\widehat{l} = \min_{l' \in [L]} \left\| \widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_{l'} \right\|_F. \tag{3.8}$$

Then, we have $\mathbb{P}(\widehat{l}(\mathbf{Y}) \neq l) = 0$ for the minimum-distance detector $\widehat{l}(\mathbf{Y})$ as long as $\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F < \sqrt{\gamma\varepsilon}$. The goal then is to relate ε to $\mathbb{P}(\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F \geq \sqrt{\gamma\varepsilon})$ and $\mathbb{P}(\widehat{l}(\mathbf{Y}) \neq l)$ using Fano's inequality [59]:

$$(1 - \mathbb{P}(\widehat{l}(\mathbf{Y}) \neq l)) \log_2 L - 1 \leq I(\mathbf{Y}; l), \tag{3.9}$$

where $I(\mathbf{Y}; l)$ denotes the mutual information (MI) between the observations $\mathbf{Y}$ and the dictionary $\mathbf{D}_l$. Notice that the smaller α_L is in (3.7), the smaller $I(\mathbf{Y}; l)$ will be in (3.9). Unfortunately, explicitly evaluating $I(\mathbf{Y}; l)$ is a challenging task in our setup because the underlying distributions are mixture of distributions. Similar to [1], we will instead resort to upper bounding $I(\mathbf{Y}; l)$ by conditioning it on some side information $\mathbf{T}(\mathbf{X})$ that will make the observations $\mathbf{Y}$ conditionally multivariate Gaussian (in particular, from [1, Lemma A.1], it follows that $I(\mathbf{Y}; l) \leq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$).[5] We will in particular focus on two types of side information: $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ and $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$. A lower bound on the minimax risk in this setting depends not only on problem parameters

[5]Instead of upper bounding $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$, similar results can be derived by using Fano's inequality for the conditional probability of error, $\mathbb{P}(\widehat{l}(\mathbf{Y}) \neq l|\mathbf{T}(\mathbf{X}))$[78, Theorem 2].

such as the number of observations N, noise variance σ^2, dimensions $\{m_k\}_{k=1}^K$ and $\{p_k\}_{k=1}^K$ of the true KS dictionary, neighborhood radius r, and coefficient covariance $\mathbf{\Sigma}_x$, but also on the structure of the constructed class $\mathcal{D}_L$ [75]. Note that our approach is applicable to the global KS-DL problem, since the minimax lower bounds that are obtained for any $\mathbf{D}^0 \in \mathcal{X}(\mathbf{D}^*, r)$ are also trivially lower bounds for $\mathbf{D}^0 \in \mathcal{D}_{KS}$.

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

3.2.2 Coefficient Distribution

By making different assumptions on coefficient distributions, we can specialize our lower bounds to specific cases. To facilitate comparisons with prior work, we adopt somewhat similar coefficient distributions as in the unstructured case [1]. First, we consider any coefficient distribution and only assume that the coefficient covariance matrix exists. We then specialize our analysis to sparse coefficient vectors and, by adding additional conditions on the reference dictionary $\mathbf{D}^*$, we obtain a tighter lower bound for the minimax risk for some SNR regimes.

General Coefficients

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

Sparse Coefficients

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

$\text{supp}(\mathbf{x})$:[6]

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

$$\mathbb{P}(\text{supp}(\mathbf{x}) = \mathcal{S}) = \frac{1}{\binom{p}{s}}, \quad \text{for any } \mathcal{S} \in \mathcal{E}_1. \tag{3.10}$$

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

$$\mathbb{P}(\text{supp}(\mathbf{x}) = \mathcal{S}) = \frac{1}{\prod_{k\in[K]} \binom{p_k}{s_k}}, \quad \text{for any } \mathcal{S} \in \mathcal{E}_2. \tag{3.11}$$

In other words, separable sparsity requires non-zero coefficients to be grouped in blocks. This model arises in the case of processing of images and video sequences [28].

Remark 3.1. If $\underline{\mathbf{X}}$ follows the separable sparsity model with sparsity $(s_1, \ldots, s_K)$, then the columns of the mode-k matrix $\mathbf{Y}_{(k)}$ of $\underline{\mathbf{Y}}$ have s_k-sparse representations with respect to $\mathbf{D}_k^0$, for $k \in [K]$ [28].

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

$$\mathbb{E}_x\{\mathbf{x}_\mathcal{S}\mathbf{x}_\mathcal{S}^T|\mathcal{S}\} = \sigma_a^2\mathbf{I}_s. \tag{3.12}$$

Any $\mathbf{x}$ with sparsity model (3.10) or (3.11) and nonzero entries satisfying (3.12) has covariance matrix

$$\mathbf{\Sigma}_x = \frac{s}{p}\sigma_a^2\mathbf{I}_p. \tag{3.13}$$

[6]These sparse coefficient models were also briefly discussed in 2.3.2.

3.3 Lower Bound for General Distribution

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

Theorem 3.1. *Consider a KS-DL problem with N i.i.d. observations generated according to model* (3.1)*. Suppose the true dictionary satisfies* (3.5) *for some r and fixed reference dictionary $\mathbf{D}^*$ satisfying* (3.4)*. Then for any coefficient distribution with mean zero and covariance $\mathbf{\Sigma}_x$, we have the following lower bound on ε^*:*

$$\varepsilon^* \geq \frac{t}{4} \min \left\{ p, \frac{r^2}{2K}, \frac{\sigma^2}{4NK\|\mathbf{\Sigma}_x\|_2} \left(c_1 \left(\sum_{k \in [K]} (m_k - 1) p_k \right) - \frac{K}{2} \log_2 2K - 2 \right) \right\}, \quad (3.14)$$

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

The implications of Theorem 3.1 are examined in Section 3.6.

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

$$\|\mathbf{D}_l - \mathbf{D}'_l\|_F \geq 2\sqrt{2\varepsilon}, \text{ for } l \neq l'. \quad (3.15)$$

In this case, if a dictionary $\mathbf{D}_l \in \mathcal{D}_L$ is selected uniformly at random from $\mathcal{D}_L$, then conditioned on side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, the observations under this dictionary follow a multivariate Gaussian distribution. We can therefore upper bound the conditional MI by approximating the upper bound for KL-divergence of multivariate Gaussian distributions. This bound depends on parameters $\varepsilon, N, \{m_k\}_{k=1}^K, \{p_k\}_{k=1}^K, \mathbf{\Sigma}_x, s, r, K$, and σ^2.

Assuming (3.15) holds for $\mathcal{D}_L$, if there exists an estimator achieving the minimax risk $\varepsilon^* \leq \varepsilon$ and the recovered dictionary $\widehat{\mathbf{D}}(\mathbf{Y})$ satisfies $\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F < \sqrt{2\varepsilon}$, the minimum distance detector can recover $\mathbf{D}_l$. Then, using the Markov inequality and since ε^* is bounded, the probability of error $\mathbb{P}(\widehat{\mathbf{D}}(\mathbf{Y}) \neq \mathbf{D}_l) \leq \mathbb{P}(\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F \geq \sqrt{2\varepsilon})$ can be

upper bounded by $\frac{1}{2}$. Further, according to (3.9), the lower bound for the conditional MI can be obtained using Fano's inequality [1]. The lower bound is a function of L only. Finally, using the obtained bounds for the conditional MI, we derive a lower bound for the minimax risk ε^*.

Remark 3.2. We use the constraint in (3.15) in Theorem 3.1 for simplicity: the number $2\sqrt{2}$ can be replaced with any arbitrary $\gamma > 0$.

The complete technical proof of Theorem 3.1 relies on the following lemmas, which are formally proved in the appendix. Although the similarity of our model to that of Jung et al. [1] suggests that our proof should be a simple extension of their proof of Theorem 1, the construction for KS dictionaries is more complex and its analysis requires a different approach. One exception is Lemma 3.3 [1, Lemma 8], which connects a lower bound on the Frobenius norms of pairwise differences in the construction to a lower bound on the conditional MI used in Fano's inequality [59].

Lemma 3.1. *Let $\alpha > 0$ and $\beta > 0$. Let $\{\mathbf{A}_l \in \mathbb{R}^{m\times p} : l \in [L]\}$ be a set of L matrices where each $\mathbf{A}_l$ contains $m \times p$ independent and identically distributed random variables taking values $\pm\alpha$ uniformly. Then we have the following inequality:*

$$\mathbb{P}\left(\exists(l,l') \in [L]\times[L], l \neq l' : |\langle \mathbf{A}_l, \mathbf{A}_{l'}\rangle| \geq \beta\right) \leq 2L^2 \exp\left(-\frac{\beta^2}{4\alpha^4 mp}\right). \tag{3.16}$$

Lemma 3.2. *Consider the generative model in* (3.1). *Fix $r > 0$ and a reference dictionary $\mathbf{D}^*$ satisfying* (3.4). *Then there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}^*, r)$ of cardinality $L = 2^{\lfloor c_1(\sum_{k\in[K]}(m_k-1)p_k) - \frac{K}{2}\log_2(2K)\rfloor}$ such that for any $0 < t < 1$, any $0 < c_1 < \frac{t^2}{8\log 2}$, any $\varepsilon' > 0$ satisfying*

$$\varepsilon' < r^2 \min\left\{1, \frac{r^2}{2Kp}\right\}, \tag{3.17}$$

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

$$\frac{2p}{r^2}(1-t)\varepsilon' \leq \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \leq \frac{4Kp}{r^2}\varepsilon'. \tag{3.18}$$

Furthermore, if $\mathbf{X}$ is drawn from a distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_x$

and conditioning on side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$*, we have*

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq \frac{2NKp\|\mathbf{\Sigma}_x\|_2}{r^2\sigma^2}\varepsilon'. \tag{3.19}$$

Lemma 3.3 (Lemma 8 [1])**.** *Consider the generative model in* (3.1) *and suppose the minimax risk* ε^* *satisfies* $\varepsilon^* \leq \varepsilon$ *for some* $\varepsilon > 0$*. If there exists a finite set* $\mathcal{D}_L \subseteq \mathcal{D}$ *with* L *dictionaries satisfying*

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \geq 8\varepsilon \tag{3.20}$$

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

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \geq \frac{1}{2}\log_2(L) - 1. \tag{3.21}$$

Proof of Lemma 3.3. The proof of Lemma 3.3 is identical to the proof of Lemma 8 in Jung et al. [1]. □

Proof of Theorem 3.1. According to Lemma 3.2, for any ε' satisfying (3.17), there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}^*, r)$ of cardinality $L = 2^{\lfloor c_1(\sum_{k\in[K]}(m_k-1)p_k) - \frac{K}{2}\log_2(2K)\rfloor}$ that satisfies (3.19) for any $0 < t' < 1$ and any $c_1 < \dfrac{t'}{8\log 2}$. Let $t = 1 - t'$. If there exists an estimator with worst-case MSE satisfying $\varepsilon^* \leq \dfrac{2tp}{8}\min\left\{1, \dfrac{r^2}{2Kp}\right\}$ then, according to Lemma 3.3, if we set $\frac{2tp}{r^2}\varepsilon' = 8\varepsilon^*$, (3.20) is satisfied for $\mathcal{D}_L$ and (3.21) holds. Combining (3.19) and (3.21) we get

$$\frac{1}{2}\log_2(L) - 1 \leq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq \frac{16NKp\|\mathbf{\Sigma}_x\|_2}{c_2 r^2\sigma^2}\varepsilon^*, \tag{3.22}$$

where $c_2 = \dfrac{2tp}{r^2}$. We can write (3.22) as

$$\varepsilon^* \geq \frac{t\sigma^2}{16NK\|\mathbf{\Sigma}_x\|_2}\left(c_1\left(\sum_{k\in[K]}(m_k-1)p_k\right) - \frac{K}{2}\log_2 2K - 2\right). \tag{3.23}$$

□

3.4 Lower Bound for Sparse Distributions

We now turn our attention to the case of sparse coefficients and obtain lower bounds for the corresponding minimax risk. We first state a corollary of Theorem 3.1 for sparse coefficients, corresponding to $\mathbf{T}(\mathbf{X}) = \mathbf{X}$.

Corollary. *Consider a KS-DL problem with N i.i.d. observations generated according to model* (3.1)*. Suppose the true dictionary satisfies* (3.5) *for some r and fixed reference dictionary $\mathbf{D}_0$ satisfying* (3.4)*. If the random coefficient vector $\mathbf{x}$ is selected according to* (3.10) *or* (3.11)*, we have the following lower bound on ε^*:*

$$\varepsilon^* \geq \frac{t}{4} \min \left\{ p, \frac{r^2}{2K}, \frac{\sigma^2 p}{4NKs\sigma_a^2} \left(c_1 \left(\sum_{k \in [K]} (m_k - 1) p_k \right) - \frac{K}{2} \log_2 2K - 2 \right) \right\}, \tag{3.24}$$

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

This result is a direct consequence of Theorem 3.1, obtained by substituting the covariance matrix of sparse coefficients given in (3.13) into (3.14).

3.4.1 Sparse Gaussian Coefficients

In this section, we make an additional assumption on the coefficient vectors generated according to (3.10) and assume non-zero elements of the vectors follow a Gaussian distribution. By additionally assuming the non-zero entries of $\mathbf{x}$ are i.i.d. Gaussian distributed, we can write $\mathbf{x}_{\mathcal{S}}$ as

$$\mathbf{x}_{\mathcal{S}} \sim \mathcal{N}(\mathbf{0}, \sigma_a^2 \mathbf{I}_s). \tag{3.25}$$

As a result, conditioned on side information $\mathbf{T}(\mathbf{x}_n) = \text{supp}(\mathbf{x}_n)$, observations $\mathbf{y}_n$ follow a multivariate Gaussian distribution.

We now provide a lower bound on the minimax risk in the case of coefficients selected according to (3.10) and (3.25).

Theorem 3.2. *Consider a KS-DL problem with N i.i.d. observations generated according to model* (3.1)*. Suppose the true dictionary satisfies* (3.5) *for some r and fixed reference dictionary satisfying* (3.4)*. If the reference coordinate dictionaries $\{\mathbf{D}_{0,k}, k \in [K]\}$ satisfy* $\mathsf{RIP}(s, \frac{1}{2})$ *and the random coefficient vector $\mathbf{x}$ is selected according to* (3.10) *and* (3.25)*, we have the following lower bound on ε^*:*

$$\varepsilon^* \geq \frac{t}{4} \min \left\{ \frac{p}{s}, \frac{r^2}{2K}, \frac{\sigma^4 p}{36(3^{4K}) N s^2 \sigma_a^4} \left(c_1 \left(\sum_{k \in [K]} (m_k - 1) p_k \right) - \frac{1}{2} \log_2 2K - 2 \right) \right\}, \tag{3.26}$$

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

Note that in Theorem 3.2, $\mathbf{D}$ (or its coordinate dictionaries) need not satisfy the RIP condition. Rather, the RIP is only needed for the coordinate reference dictionaries, $\{\mathbf{D}_{0,k}, k \in [K]\}$, which is a significantly weaker (and possibly trivial to satisfy) condition. We state a variation of Lemma 3.2 necessary for the proof of Theorem 3.2 — the proof is provided in the appendix.

Lemma 3.4. *Consider the generative model in* (3.1)*. Fix $r > 0$ and reference dictionary $\mathbf{D}^*$ satisfying* (3.4)*. Then, there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}^*, r)$ of cardinality $L = 2^{\lfloor c_1 (\sum_{k \in [K]} (m_k - 1) p_k) - \frac{1}{2} \log_2 (2K) \rfloor}$ such that for any $0 < t < 1$, any $0 < c_1 < \frac{t^2}{8 \log 2}$, any $\varepsilon' > 0$ satisfying*

$$0 < \varepsilon' \leq r^2 \min \left\{ \frac{1}{s}, \frac{r^2}{2Kp} \right\}, \tag{3.27}$$

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

$$\frac{2p}{r^2}(1-t)\varepsilon' \leq \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \leq \frac{4Kp}{r^2}\varepsilon'. \tag{3.28}$$

Furthermore, assuming the reference coordinate dictionaries $\{\mathbf{D}_k^*, k \in [K]\}$ satisfy

$\mathsf{RIP}(s, \frac{1}{2})$, *the coefficient matrix* $\mathbf{X}$ *is selected according to* (3.10) *and* (3.25), *and considering side information* $\mathbf{T}(\mathbf{X}) = \mathrm{supp}(\mathbf{X})$, *we have:*

$$I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq 36(3^{4K}) \left(\frac{\sigma_a}{\sigma}\right)^4 \frac{Ns^2}{r^2}\varepsilon'. \tag{3.29}$$

Proof of Theorem 3.2. According to Lemma 3.4, for any ε' satisfying (3.27), there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}^*, r)$ of cardinality $L = 2^{\lfloor c_1(\sum_{k\in[K]}(m_k-1)p_k) - \frac{K}{2}\log_2(2K)\rfloor}$ that satisfies (3.29) for any $0 < t' < 1$ and any $c_1 < \frac{t'}{8\log 2}$. Denoting $t = 1 - t'$ and provided there exists an estimator with worst case MSE satisfying $\varepsilon^* \leq \frac{tp}{4}\min\{\frac{1}{s}, \frac{r^2}{2Kp}\}$, if we set $\frac{2tp}{r^2}\varepsilon' = 8\varepsilon^*$, (3.20) is satisfied for $\mathcal{D}_L$ and (3.21) holds. Consequently,

$$\frac{1}{2}\log_2(L) - 1 \leq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) \leq \frac{36(3^{4K})}{c_2}\left(\frac{\sigma_a}{\sigma}\right)^4 \frac{Ns^2}{r^2}\varepsilon^*, \tag{3.30}$$

where $c_2 = \frac{p(1-t)}{4r^2}$. We can write (3.30) as

$$\varepsilon^* \geq \left(\frac{\sigma}{\sigma_a}\right)^4 \frac{tp\left(c_1\left(\sum_{k\in[K]}(m_k-1)p_k\right) - \frac{K}{2}\log_2 2K - 2\right)}{144(3^{4K})Ns^2}. \tag{3.31}$$

□

Focusing on the case where the coefficients follow the separable sparsity model, the next theorem provides a lower bound on the minimax risk for coefficients selected according to (3.11) and (3.25).

Theorem 3.3. *Consider a KS-DL problem with* N *i.i.d. observations generated according to model* (3.1). *Suppose the true dictionary satisfies* (3.5) *for some* r *and fixed reference dictionary satisfying* (3.4). *If the reference coordinate dictionaries* $\{\mathbf{D}_k^*, k \in [K]\}$ *satisfy* $\mathsf{RIP}(s, \frac{1}{2})$ *and the random coefficient vector* $\mathbf{x}$ *is selected according to* (3.11) *and* (3.25), *we have the following lower bound on* ε^**:*

$$\varepsilon^* \geq \frac{t}{4}\min\left\{p, \frac{r^2}{2K}, \frac{\sigma^4 p}{36(3^{4K})Ns^2\sigma_a^4}\left(c_1\left(\sum_{k\in[K]}(m_k-1)p_k\right) - \frac{1}{2}\log_2 2K - 2\right)\right\}, \tag{3.32}$$

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

We state a variation of Lemma 3.4 necessary for the proof of Theorem 3.3. The proof of the lemma is provided in the appendix.

Lemma 3.5. *Consider the generative model in* (3.1). *Fix* $r > 0$ *and reference dictionary* $\mathbf{D}^*$ *satisfying* (3.4). *Then, there exists a set of dictionaries* $\mathcal{D}_L \subseteq \mathcal{D}$ *of cardinality* $L = 2^{\lfloor c_1 (\sum_{k\in[K]}(m_k-1)p_k) - \frac{K}{2}\log_2(2K)\rfloor}$ *such that for any* $0 < t < 1$, *any* $0 < c_1 < \frac{t^2}{8\log 2}$, *any* $\varepsilon' > 0$ *satisfying*

$$0 < \varepsilon' \leq r^2 \min\left\{1, \frac{r^2}{2Kp}\right\}, \tag{3.33}$$

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

$$\frac{2p}{r^2}(1-t)\varepsilon' \leq \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \leq \frac{4Kp}{r^2}\varepsilon'. \tag{3.34}$$

Furthermore, assuming the coefficient matrix $\mathbf{X}$ *is selected according to* (3.11) *and* (3.25), *the reference coordinate dictionaries* $\{\mathbf{D}_k^*, k \in [K]\}$ *satisfy* $\mathsf{RIP}(s_k, \frac{1}{2})$, *and considering side information* $\mathbf{T}(\mathbf{X}) = \mathrm{supp}(\mathbf{X})$, *we have:*

$$I(\mathbf{Y}; l | \mathbf{T}(\mathbf{X})) \leq 36(3^{4K})\left(\frac{\sigma_a}{\sigma}\right)^4 \frac{Ns^2}{r^2}\varepsilon'. \tag{3.35}$$

Proof of Theorem 3.3. The proof of Theorem 3.3 follows similar steps as the proof of Theorem 3.2. The dissimilarity arises in the condition in (3.33) for Lemma 3.5, which is different from the condition in (3.27) for Lemma 3.4. This changes the range for the minimax risk ε^* in which the lower bound in (3.31) holds. □

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

3.5 Partial Converse

In the previous sections, we provided lower bounds on the minimax risk for various coefficient vector distributions and corresponding side information. We now study a

special case of the problem and introduce an algorithm that achieves the lower bound in Corollary 3.4 (order-wise) for 2nd-order tensors. This demonstrates that our obtained lower bounds are tight in some cases.

Theorem 3.4. *Consider a DL problem with N i.i.d observations according to model* (3.1) *for $K = 2$ and let the true dictionary satisfy* (3.5) *for $\mathbf{D}^* = \mathbf{I}_p$ and some $r > 0$. Further, assume the random coefficient vector $\mathbf{x}$ is selected according to* (3.10)*, $\mathbf{x} \in \{-1, 0, 1\}^p$, where the probabilities of the nonzero entries of $\mathbf{x}$ are arbitrary. Next, assume noise standard deviation σ and express the KS dictionary as*

$$\mathbf{D}^0 = (\mathbf{I}_{p_1} + \mathbf{\Delta}_1) \otimes (\mathbf{I}_{p_2} + \mathbf{\Delta}_2), \tag{3.36}$$

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

$$\begin{aligned} &r_1\sqrt{p_2} + r_2\sqrt{p_1} + r_1 r_2 \leq r, \qquad (r_1 + r_2 + r_1 r_2)\sqrt{s} \leq 0.1 \\ &\max\left\{\frac{r_1^2}{p_2}, \frac{r_2^2}{p_1}\right\} \leq \frac{1}{3N}, \qquad \sigma \leq 0.4, \end{aligned} \tag{3.37}$$

there exists a DL scheme whose MSE satisfies

$$\mathbb{E}_{\mathbf{Y}}\left\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}^0\|_F^2\right\} \leq \frac{8p}{N}\left(\frac{p_1 m_1 + p_2 m_2}{m\,\mathrm{SNR}} + 3(p_1 + p_2)\right) + 8p\exp\left(-\frac{0.08pN}{\sigma^2}\right), \tag{3.38}$$

for any $\mathbf{D}^0 \in \mathcal{X}(\mathbf{D}^*, r)$ that satisfies (3.36) *.*

To prove Theorem 3.4, we first introduce an algorithm to learn a KS dictionary for 2nd-order tensor data. Then, we analyze the performance of the proposed algorithm and obtain an upper bound for the MSE in the proof of Theorem 3.4, which is provided in the appendix.[7] Finally, we provide numerical experiments to validate our obtained results.

[7]Theorem 3.4 also implicitly uses the assumption that $\max\{p_1, p_2\} \leq N$.

3.5.1 KS Dictionary Learning Algorithm

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

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

$$\widehat{\mathbf{x}}_n = (\widehat{x}_{n,1}, \ldots, \widehat{x}_{n,p})^\top,\ \widehat{x}_{n,l} = \begin{cases} 1 & \text{if } y_{n,l} > 0.5, \\ -1 & \text{if } y_{n,l} < -0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{3.39}$$

Dictionary Estimate: Denoting $\mathbf{A} \triangleq \mathbf{I}_{p_1} + \mathbf{\Delta}_1$ and $\mathbf{B} \triangleq \mathbf{I}_{p_2} + \mathbf{\Delta}_2$, we can write $\mathbf{D}^0 = \mathbf{A} \otimes \mathbf{B}$. We estimate the columns of $\mathbf{A}$ and $\mathbf{B}$ separately. To learn $\mathbf{A}$, we take advantage of the Kronecker structure of the dictionary and divide each observation $\mathbf{y}_n \in \mathbb{R}^{p_1 p_2}$ into p_2 observations $\mathbf{y}'_{(n,j)} \in \mathbb{R}^{p_1}$:

$$\mathbf{y}'_{(n,j)} = \{y_{n,p_2 i+j}\}_{i=0}^{p_1-1},\ j \in [p_2],\ n \in [N]. \tag{3.40}$$

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

$$\begin{aligned} \mathbf{x}'_{(n,j)} &= \{x_{n,p_2 i+j}\}_{i=0}^{p_1-1}, \\ \widehat{\mathbf{x}}'_{(n,j)} &= \{\widehat{x}_{n,p_2 i+j}\}_{i=0}^{p_1-1},\ j \in [p_2],\ n \in [N]. \end{aligned} \tag{3.41}$$

Similarly, we define new noise vectors:

$$\mathbf{w}'_{(n,j)} = \{w_{n,p_2 i+j}\}_{i=0}^{p_1-1},\ j \in [p_2],\ n \in [N]. \tag{3.42}$$

To motivate the estimation rule for the columns of $\mathbf{A}$, let us consider the original DL formulation, $\mathbf{y}_n = \mathbf{D}^0 \mathbf{x}_n + \mathbf{w}_n$, which we can rewrite as $\mathbf{y}_n = \mathbf{x}_{n,l} \mathbf{d}_l^0 + \sum_{i \neq l} \mathbf{x}_{n,i} \mathbf{d}_i^0 +$

$\mathbf{w}_n$. Multiplying both sides of the equation by $\mathbf{x}_{n,l}$ and summing up over all training data, we get $\sum_{n=1}^{N} \mathbf{x}_{n,l}\mathbf{y}_n = \sum_{n=1}^{N}(\mathbf{x}_{n,l}^2\mathbf{d}_l^0 + \sum_{i\neq l} \mathbf{x}_{n,l}\mathbf{x}_{n,i}\mathbf{d}_i^0 + \mathbf{x}_{n,l}\mathbf{w}_n)$. Using the facts $\mathbb{E}_{\mathbf{x}}\{\mathbf{x}_{n,l}^2\} = \frac{s}{p}$, $\mathbb{E}_{\mathbf{x}}\{\mathbf{x}_{n,l}\mathbf{x}_{n,i}\} = 0$ for $l \neq i$, and $\mathbb{E}_{\mathbf{x},\mathbf{w}}\{\mathbf{x}_{n,l}\mathbf{w}_n\} = 0$, we get the following approximation, $\mathbf{d}_l^0 \approx \frac{p}{Ns}\sum_{n=1}^{N} x_{n,l}\mathbf{y}_n$.[8] This suggests that for estimating the columns of $\mathbf{A}$, we can utilize the following equation:

$$\widetilde{\mathbf{a}}_l = \frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2} x'_{(k,j),l}\mathbf{y}'_{(n,j)},\ l \in [p_1]. \tag{3.43}$$

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

$$\mathbf{y}''_{(n,j)} = \left\{y_{n,i+p_1(j-1)}\right\}_{i=1}^{p_2},\ j \in [p_1],\ n \in [N]. \tag{3.44}$$

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

$$\begin{aligned}\mathbf{x}''_{(n,j)} &= \left\{x_{k,i+p_1(j-1)}\right\}_{i=0}^{p_1-1},\\ \widehat{\mathbf{x}}''_{(n,j)} &= \left\{\widehat{x}_{n,i+p_1(j-1)}\right\}_{i=0}^{p_1-1},\ j \in [p_1],\ n \in [N].\end{aligned} \tag{3.45}$$

Similarly, we define new noise vectors:

$$\mathbf{w}''_{(n,j)} = \left\{w_{n,i+p_1(j-1)}\right\}_{i=1}^{p_2},\ j \in [p_1],\ n \in [N]. \tag{3.46}$$

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

$$\widetilde{\mathbf{b}}_l = \frac{p_2}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_1} x''_{(n,j),l}\mathbf{y}''_{(n,j)},\ l \in [p_2]. \tag{3.47}$$

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

The final estimate for the recovered dictionary is

$$\begin{aligned}
&\widehat{\mathbf{D}} = \widehat{\mathbf{A}} \otimes \widehat{\mathbf{B}}, \\
&\widehat{\mathbf{A}} = (\widehat{\mathbf{a}}_1, \ldots, \widehat{\mathbf{a}}_{p_1}), \quad \widehat{\mathbf{a}}_l = P_{\mathcal{B}_1}(\widetilde{\mathbf{a}}_l), \\
&\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_{p_2}), \quad \widehat{\mathbf{b}}_l = P_{\mathcal{B}_1}(\widetilde{\mathbf{b}}_l),
\end{aligned} \tag{3.48}$$

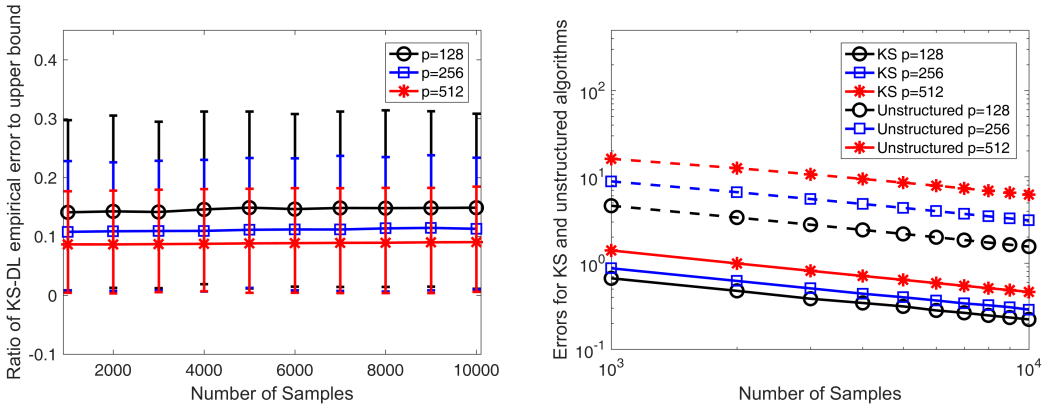
where the projection on the closed unit ball ensures that $\|\widehat{\mathbf{a}}_l\|_2 \leq 1$ and $\|\widehat{\mathbf{b}}_l\|_2 \leq 1$. Note that although projection onto the closed unit ball does not ensure the columns of $\widehat{\mathbf{D}}$ to have unit norms, our analysis only imposes this condition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

Remark 3.3. In addition to the heuristics following (3.42), the exact update rules for $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ in (3.43) and (3.47) require some additional perturbation analysis. To see this for the case of $\widetilde{\mathbf{A}}$, notice that (3.43) follows from writing $\mathbf{A} \otimes \mathbf{B}$ as $\mathbf{A} \otimes (\mathbf{I}_{p_2} + \boldsymbol{\Delta}_2)$, rearranging each $\mathbf{y}_n$ and $(\mathbf{A} \otimes \mathbf{I}_{p_2})\mathbf{x}_n$ into $\mathbf{y}'_{(n,j)}$'s and $\mathbf{A}\mathbf{x}'_{(n,j)}$'s, and using them to update $\widetilde{\mathbf{A}}$. In this case, we treat $(\mathbf{A} \otimes \boldsymbol{\Delta}_2)\mathbf{x}_n$ as a perturbation term in our analysis. A similar perturbation term appears in the case of the update rule for $\widetilde{\mathbf{B}}$. The analysis for dealing with these perturbation terms is provided in the appendix.

3.5.2 Empirical Comparison to Upper Bound

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

Figure 3.1a shows the ratio of the empirical error of the proposed KS-DL algorithm in Section 3.5.1 to the obtained upper bound in Theorem 3.4 for 50 Monte Carlo experiments. This ratio is plotted as a function of the sample size for three choices of the number of columns p: 128, 256, and 512. The experiment shows that the ratio is approximately constant as a function of sample size, verifying the theoretical result that the estimator meets the minimax bound in terms of error scaling as a function of sample size. Figure 3.1b shows the performance of our KS-DL algorithm in relation to the unstructured DL algorithm provided in [1]. It is evident that the error of our



(a) (b)

Figure 3.1: Performance summary of KS-DL algorithm for $p = \{128, 256, 512\}$, $s = 5$ and $r = 0.1$. (a) plots the ratio of the empirical error of our KS-DL algorithm to the obtained error upper bound along with error bars for generated square KS dictionaries, and (b) shows the performance of our KS-DL algorithm (solid lines) compared to the unstructured learning algorithm proposed in [1] (dashed lines).

algorithm is significantly less than that for the unstructured algorithm for all three choices of p. This verifies that taking the structure of the data into consideration can indeed lead to lower dictionary identification error.

3.6 Discussion

We now discuss some of the implications of our results. Table 3.1 summarizes the lower bounds on the minimax rates from previous papers and this chapter. The bounds are given in terms of the number of component dictionaries K, the dictionary size parameters (m_k's and p_k's), the coefficient distribution parameters, the number of samples N, and SNR, which is defined as

$$\text{SNR} = \frac{\mathbb{E}_{\mathbf{x}}\left\{\|\mathbf{x}\|_2^2\right\}}{\mathbb{E}_{\mathbf{w}}\left\{\|\mathbf{w}\|_2^2\right\}} = \frac{\text{Tr}(\mathbf{\Sigma}_x)}{m\sigma^2}. \tag{3.49}$$

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

Comparison of minimax lower bounds for unstructured and KS DL: Compared to the results for the unstructured DL problem [1], we are able to decrease the lower bound for various coefficient distributions by reducing the scaling $\Omega(mp)$ to $\Omega(\sum_{k\in[K]} m_k p_k)$ for KS dictionaries. This is intuitively pleasing since the minimax

Table 3.1: Order-wise lower bounds on the minimax risk for various coefficient distributions

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

lower bound has a linear relationship with the number of degrees of freedom of the KS dictionary, which is $\sum_{k\in[K]} m_k p_k$.

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

Looking at the results for the general coefficient model in the first row of Table 3.1, the lower bound for any arbitrary zero-mean random coefficient vector distribution with covariance $\mathbf{\Sigma}_x$ implies an inverse relationship between the minimax risk and SNR due to the fact that $\|\mathbf{\Sigma}_x\|_2 \leq \mathrm{Tr}(\mathbf{\Sigma}_x)$.

Comparison of general sparse and Gaussian sparse coefficient distributions: Proceeding to the sparse coefficient vector model in the second row of Table 3.1, by replacing $\mathbf{\Sigma}_x$ with the expression in (3.13) in the minimax lower bound for the general coefficient distribution, we obtain the second lower bound given in (3.24). Recall that for s-sparse coefficient vectors,

$$\mathrm{SNR} = \frac{s\sigma_a^2}{m\sigma^2}. \tag{3.50}$$

Using this definition of SNR in (3.24), we observe a seemingly counter-intuitive increase in the MSE of order $\Omega\left(p/s\right)$ in the lower bound in comparison to the general coefficient model. However, this increase is due to the fact that we do not require coefficient

vectors to have constant energy; because of this, SNR decreases for s-sparse coefficient vectors.

Next, looking at the third row of Table 3.1, by restricting the class of sparse coefficient vector distributions to the case where non-zero elements of the coefficient vector follow a Gaussian distribution according to (3.25), we obtain a minimax lower bound that involves less side information than the prior two cases. However, we do make the assumption in this case that reference coordinate dictionaries satisfy $\mathsf{RIP}(s, \frac{1}{2})$. This additional assumption has two implications: (1) it introduces the factor of $1/3^{4K}$ in the minimax lower bound, and (2) it imposes the following condition on the sparsity for the "random sparsity" model: $s \leq \min_{k\in[K]}\{p_k\}$. Nonetheless, considering sparse-Gaussian coefficient vectors, we obtain a minimax lower bound that is tighter than the previous bound for some SNR values. Specifically, in order to compare bounds obtained in (3.24) and (3.26) for sparse and sparse-Gaussian coefficient vector distributions, we fix K. Then in high SNR regimes, i.e., SNR $= \Omega(1/m)$, the lower bound in (3.24) is tighter, while (3.26) results in a tighter lower bound in low SNR regimes, i.e., SNR $= \mathcal{O}(1/m)$, which correspond to low sparsity settings.

Comparison of random and separable sparse coefficient models: We now focus on our results for the two sparsity pattern models, namely, random sparsity and separable sparsity, for the case of sparse-Gaussian coefficient vector distribution. These results, which are reported in (3.26) and (3.32), are almost identical to each other, except for the first term in the minimization. In order to understand the settings in which the separable sparsity model in (3.11)—which is clearly more restrictive than the random sparsity model in (3.10)—turns out to be more advantageous, we select the neighborhood radius r to be of order $\mathcal{O}(\sqrt{p})$; since we are dealing with dictionaries that lie on the surface of a sphere with radius $\sqrt{p}$, this effectively ensures $\mathcal{X}(\mathbf{D}^*, r) \approx \mathcal{D}$. In this case, it can be seen from (3.26) and (3.32) that if $s = \Omega(K)$ then the separable sparsity model gives a better minimax lower bound. On the other hand, the random sparsity model should be considered for the case of $s = \mathcal{O}(K)$ because of the less restrictive nature of this model.

Achievability of our minimax lower bounds for learning KS dictionaries:

To this end, we provided a simple KS-DL algorithm in Section 3.5 for the special scenario of 2-dimensional tensors and analyzed the corresponding MSE, $\mathbb{E}_{\mathbf{Y}}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}^0\|_F^2\}$. In terms of scaling, the upper bound obtained for the MSE in Theorem 3.4 matches the lower bound in Corollary 3.4 provided $p_1 + p_2 < \frac{m_1 p_1 + m_2 p_2}{m\,\mathrm{SNR}}$ holds. This result suggests that more general KS-DL algorithms may be developed to achieve the lower bounds reported in this chapter.

3.7 Conclusion

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

3.8 Appendix

3.8.1 Proof of Lemma 3.1

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

$$f(\mathbf{a}_l^\top, \mathbf{a}_{l'}^\top) \triangleq |\langle \mathbf{A}_l, \mathbf{A}_{l'} \rangle| = |\langle \mathbf{a}_l, \mathbf{a}_{l'} \rangle| \,. \tag{3.51}$$

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

$$\sup_{\widetilde{\mathbf{a}} \sim \widetilde{\mathbf{a}}'} \left| f(\widetilde{\mathbf{a}}) - f(\widetilde{\mathbf{a}}') \right| = (\alpha - (-\alpha))\alpha = 2\alpha^2. \tag{3.52}$$

Hence, according to McDiarmid's inequality [79], for all $\beta > 0$, we have

$$\begin{aligned} \mathbb{P}\left(|\langle \mathbf{A}_l, \mathbf{A}_{l'} \rangle| \geq \beta \right) &\leq 2 \exp\left(\frac{-2\beta^2}{\sum_{i=1}^{2mp} (2\alpha^2)^2} \right) \\ &= 2 \exp\left(-\frac{\beta^2}{4\alpha^4 mp} \right). \end{aligned} \tag{3.53}$$

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

$$\begin{aligned} &\mathbb{P}\left(\exists (l, l') \in [L] \times [L], l \neq l' : |\langle \mathbf{A}_l, \mathbf{A}_{l'} \rangle| \geq \beta \right) \\ &\qquad \leq 2L^2 \exp\left(-\frac{\beta^2}{4\alpha^4 mp} \right). \end{aligned} \tag{3.54}$$

3.8.2 Proof of Lemma 3.2

Fix $r > 0$ and $t \in (0, 1)$. Let $\mathbf{D}^*$ be a reference dictionary satisfying (3.4), and let $\{\mathbf{U}_{(k,j)}\}_{j=1}^{p_k} \in \mathbb{R}^{m_k \times m_k}$, $k \in [K]$, be arbitrary unitary matrices satisfying

$$\mathbf{d}_{k,j}^* = \mathbf{U}_{(k,j)} \mathbf{e}_1, \tag{3.55}$$

where $\mathbf{d}_{k,j}^*$ denotes the j-th column of $\mathbf{D}_k^*$.

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

$$L_k = 2^{\lfloor c_1(m_k-1)p_k - \frac{1}{2}\log_2 2K \rfloor} \tag{3.56}$$

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

$$\mathbf{G}_{(k,l_k)} \in \left\{ -\frac{1}{r^{1/K}\sqrt{(m_k-1)}}, \frac{1}{r^{1/K}\sqrt{(m_k-1)}} \right\}^{(m_k-1)\times p_k} \tag{3.57}$$

for $k \in [K]$ and $l_k \in [L_k]$. According to Lemma 3.1, for all $k \in [K]$ and any $\beta > 0$, the following relation is satisfied:

$$\begin{aligned} \mathbb{P}\left(\exists (l_k, l_k') \in [L_k]\times[L_k], l \neq l' : \left|\left\langle \mathbf{G}_{(k,l_k)}, \mathbf{G}_{(k,l_k')} \right\rangle\right| \geq \beta\right) \\ \leq 2L_k^2 \exp\left(-\frac{r^{4/K}(m_k-1)\beta^2}{4p_k}\right). \end{aligned} \tag{3.58}$$

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

$$\left|\left\langle \mathbf{G}_{(k,l_k)}, \mathbf{G}_{(k,l_k')} \right\rangle\right| \leq \beta, \quad k \in [K], \tag{3.59}$$

we take a union bound of (3.58) over all $k \in [K]$ and choose parameters such that the following upper bound is less than 1:

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

which is satisfied as long as the following inequality holds:

$$\log_2 L_k < \frac{r^{4/K}(m_k-1)\beta^2}{8p_k \log 2} - \frac{1}{2} - \frac{1}{2}\log_2 K. \tag{3.60}$$

Now, setting $\beta = \frac{p_k t}{r^{2/K}}$, the condition in (3.60) holds and there exists a collection of

generating matrices that satisfy:

$$\left|\left\langle \mathbf{G}_{(k,l_k)}, \mathbf{G}_{(k,l_k')} \right\rangle\right| \leq \frac{p_k t}{r^{2/K}}, \quad k \in [K], \tag{3.61}$$

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

$$c_1 < \frac{t^2}{8 \log 2}. \tag{3.62}$$

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

$$\mathbf{d}_{(k,1,l_k),j} = \mathbf{U}_{(k,j)} \begin{pmatrix} 0 \\ \\ \mathbf{g}_{(k,l_k),j} \end{pmatrix}, \quad k \in [K], \tag{3.63}$$

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

$$\mathcal{D}_1 \triangleq \left\{ \bigotimes_{k \in [K]} \mathbf{D}_{(k,1,l_k)} : l_k \in [L_k] \right\}, \tag{3.64}$$

and denoting

$$\mathcal{L} \triangleq \{(l_1, \ldots, l_K) : l_k \in [L_k]\}, \tag{3.65}$$

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

$$\mathbf{D}_{(1,l)} = \bigotimes_{k \in [K]} \mathbf{D}_{(k,1,l_k)}, \forall\ l \in [L], \tag{3.66}$$

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

$$\left\|\mathbf{d}_{(1,l),j}\right\|_2^2 \overset{(a)}{=} \prod_{k\in[K]} \left\|\mathbf{d}_{(k,1,l_k),j}\right\|_2^2 = \prod_{k\in[K]} \frac{1}{r^{2/K}} = \frac{1}{r^2}, \text{ and}$$
$$\left\|\mathbf{D}_{(1,l)}\right\|_F^2 = \frac{p}{r^2}, \tag{3.67}$$

where (a) follows from properties of the Kronecker product. From (3.63), it is evident that for all $k \in [K]$, $\mathbf{d}_{(k,0),j}$ is orthogonal to $\mathbf{d}_{(k,1,l_k),j}$ and consequently, we have

$$\left\langle \mathbf{D}_k^*, \mathbf{D}_{(k,1,l_k)} \right\rangle = 0, \ k \in [K] \tag{3.68}$$

Also,

$$\begin{aligned}
\left\langle \mathbf{D}_{(k,1,l_k)}, \mathbf{D}_{(k,1,l'_k)} \right\rangle &= \sum_{j=1}^{p_k} \left\langle \mathbf{d}_{(k,1,l_k),j}, \mathbf{d}_{(k,1,l'_k),j} \right\rangle \\
&= \sum_{j=1}^{p_k} \left\langle \mathbf{U}_{(k,j)} \begin{pmatrix} 0 \\ \\ \mathbf{g}_{(k,l_k),j} \end{pmatrix}, \mathbf{U}_{(k,j)} \begin{pmatrix} 0 \\ \\ \mathbf{g}_{(k,l'_k),j} \end{pmatrix} \right\rangle \\
&\overset{(b)}{=} \sum_{j=1}^{p_k} \left\langle \mathbf{g}_{(k,l_k),j}, \mathbf{g}_{(k,l'_k),j} \right\rangle \\
&= \left\langle \mathbf{G}_{(k,l_k)}, \mathbf{G}_{(k,l'_k)} \right\rangle,
\end{aligned} \tag{3.69}$$

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

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

$$\begin{aligned}
\left\|\mathbf{D}_{(1,l)} - \mathbf{D}_{(1,l')}\right\|_F^2 &= \left\|\mathbf{D}_{(1,l)}\right\|_F^2 + \left\|\mathbf{D}_{(1,l')}\right\|_F^2 - 2\left\langle \mathbf{D}_{(1,l)}, \mathbf{D}_{(1,l')} \right\rangle \\
&= \frac{p}{r^2} + \frac{p}{r^2} - 2\prod_{k\in[K]} \left\langle \mathbf{D}_{(k,1,l_k)}, \mathbf{D}_{(k,1,l'_k)} \right\rangle \\
&\geq 2\left(\frac{p}{r^2} - \prod_{k\in[K]} \left| \left\langle \mathbf{D}_{(k,1,l_k)}, \mathbf{D}_{(k,1,l'_k)} \right\rangle \right| \right) \\
&\overset{(c)}{=} 2\left(\frac{p}{r^2} - \prod_{k\in[K]} \left| \left\langle \mathbf{G}_{(k,l_k)}, \mathbf{G}_{(k,l'_k)} \right\rangle \right| \right)
\end{aligned}$$

$$\overset{(d)}{\geq} 2\left(\frac{p}{r^2} - \prod_{k\in[K]} \frac{p_k}{r^{2/K}} t\right)$$
$$= \frac{2p}{r^2}\left(1 - t^K\right), \tag{3.70}$$

where (c) and (d) follow from (3.69) and (3.61), respectively.

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

$$\mathcal{D}_L \triangleq \left\{ \bigotimes_{k\in[K]} \mathbf{D}_{(k,l_k)} : l_k \in [L_k] \right\} \tag{3.71}$$

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

$$\mathbf{D}_l = \bigotimes_{k\in[K]} \mathbf{D}_{(k,l_k)}, \tag{3.72}$$

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

$$\mathbf{D}_{(k,l_k)} \triangleq \eta \mathbf{D}_k^* + \nu \mathbf{D}_{(k,1,l_k)}, \quad k \in [K], \tag{3.73}$$

and

$$\eta \triangleq \sqrt{1 - \frac{\varepsilon'}{r^2}}, \quad \nu \triangleq \sqrt{\frac{r^{2/K}\varepsilon'}{r^2}} \tag{3.74}$$

for any

$$0 < \varepsilon' < \min\left\{ r^2, \frac{r^4}{2Kp} \right\}, \tag{3.75}$$

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

$$\eta^2 + \frac{\nu^2}{r^{2/K}} = 1. \tag{3.76}$$

We can expand (3.72) to facilitate the forthcoming analysis:

$$\mathbf{D}_l = \sum_{\mathbf{i}\in\{0,1\}^K} \eta^{K-\|\mathbf{i}\|_1}\nu^{\|\mathbf{i}\|_1}\Big(\bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l_k)}\Big), \tag{3.77}$$

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

$$\begin{aligned}
\|\mathbf{d}_{l,j}\|_2^2 &= \prod_{k\in[K]} \left\|\mathbf{d}_{(k,l_k),j_k}\right\|_2^2 \\
&= \prod_{k\in[K]} \left(\eta^2\|\mathbf{d}^*_{k,j_k}\|_2^2 + \nu^2\left\|\mathbf{d}_{(k,1,l_k),j_k}\right\|_2^2\right) \\
&= \prod_{k\in[K]} \left(\eta^2 + \nu^2\big(\frac{1}{r^{2/K}}\big)\right) \\
&\overset{(e)}{=} 1,
\end{aligned} \tag{3.78}$$

where (e) follows from (3.76). Then, we show that $\|\mathbf{D}_l - \mathbf{D}^*\|_F \leq r$:

$$\begin{aligned}
\|\mathbf{D}_l - \mathbf{D}^*\|_F^2 &= \Big\|\ \mathbf{D}^* - \sum_{\mathbf{i}\in\{0,1\}^K} \eta^{K-\|\mathbf{i}\|_1}\nu^{\|\mathbf{i}\|_1} \bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l_k)}\Big\|_F^2 \\
&= \Big\|\big(1-\eta^K\big)\mathbf{D}^* - \sum_{\substack{\mathbf{i}\in\{0,1\}^K \\ \|\mathbf{i}\|_1\neq 0}} \eta^{K-\|\mathbf{i}\|_1}\nu^{\|\mathbf{i}\|_1} \bigotimes_{k\in[K]} \mathbf{D}_{(k,i_k,l_k)}\Big\|_F^2 \\
&= \big(1-\eta^K\big)^2\|\mathbf{D}^*\|_F^2 + \sum_{\substack{\mathbf{i}\in\{0,1\}^K \\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)}\nu^{2\|\mathbf{i}\|_1} \prod_{k\in[K]} \left\|\mathbf{D}_{(k,i_k,l_k)}\right\|_F^2.
\end{aligned} \tag{3.79}$$

We will bound the two terms in (3.79) separately. We know

$$(1-x^n) = (1-x)(1+x+x^2+\cdots+x^{n-1}). \tag{3.80}$$

Hence, we have

$$\begin{aligned}
\left(1-\eta^K\right)^2 \|\mathbf{D}^*\|_F^2 &= \left(1-\eta^K\right)^2 p \\
&\overset{(f)}{\leq} \left(1-\eta^K\right) p \\
&\leq \left(1-\eta^{2K}\right) p \\
&\overset{(g)}{=} \left(1-\eta^2\right)\left(1+\eta^2+\cdots+\eta^{2(K-1)}\right) p \\
&= \frac{\varepsilon'}{r^2}\left(1+\eta^2+\cdots+\eta^{2(K-1)}\right) p \\
&\overset{(h)}{\leq} \frac{Kp\varepsilon'}{r^2},
\end{aligned} \tag{3.81}$$

where (f) and (h) follow from the fact that $\eta < 1$ and (g) follows from (3.80).

Similarly for the second term in (3.79),

$$\begin{aligned}
\prod_{k\in[K]} \left\|\mathbf{D}_{(k,i_k,l_k)}\right\|_F^2 &= \Bigg(\prod_{\substack{k\in[K]\\ i_k=0}} \|\mathbf{D}_k^*\|_F^2\Bigg)\Bigg(\prod_{\substack{k\in[K]\\ i_k=1}} \|\mathbf{D}_{(k,1,l_k)}\|_F^2\Bigg) \\
&= \Bigg(\prod_{\substack{k\in[K]\\ i_k=0}} p_k\Bigg)\Bigg(\prod_{\substack{k\in[K]\\ i_k=1}} \frac{p_k}{r^{2/K}}\Bigg) \\
&= \Bigg(\prod_{k\in[K]} p_k\Bigg)\left(\frac{1}{r^{2/K}}\right)^{\|\mathbf{i}\|_1}.
\end{aligned} \tag{3.82}$$

Replacing values for η and ν from (3.74) and using (3.82) and the fact that $\prod_{k\in[K]} p_k = p$, we can further reduce the second term in (3.79) to get

$$\begin{aligned}
&\sum_{\substack{\mathbf{i}\in\{0,1\}^K\\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)}\nu^{2\|\mathbf{i}\|_1} \prod_{k\in[K]} \|\mathbf{D}_{(k,i_k,l_k)}\|_F^2 \\
&\quad= p\sum_{k=0}^{K-1}\binom{K}{k}\left(1-\frac{\varepsilon'}{r^2}\right)^k\left(\frac{\varepsilon'}{r^2}\right)^{K-k} \\
&\quad= p\left(1-\left(1-\frac{\varepsilon'}{r^2}\right)^K\right) \\
&\quad\overset{(i)}{=} p\left(\frac{\varepsilon'}{r^2}\right)\left(1+\left(1-\frac{\varepsilon'}{r^2}\right)+\cdots+\left(1-\frac{\varepsilon'}{r^2}\right)^{K-1}\right) \\
&\quad\leq \frac{Kp\varepsilon'}{r^2},
\end{aligned} \tag{3.83}$$

where (i) follows from (3.80). Adding (3.81) and (3.83), we get

$$\begin{aligned}\|\mathbf{D}_l - \mathbf{D}^*\|_F^2 &\leq \varepsilon' \left(\frac{2Kp}{r^2}\right) \\ &\overset{(j)}{\leq} r^2, \end{aligned} \tag{3.84}$$

where (j) follows from the condition in (3.75). Therefore, (3.78) and (3.83) imply that $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}^*, r)$.

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

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

$$\begin{aligned}\left\| \bigotimes_{k\in\mathcal{I}_i} \mathbf{D}_{(k,1,l_k)} - \bigotimes_{k\in\mathcal{I}_i} \mathbf{D}_{(k,1,l'_k)} \right\|_F^2 &\overset{(k)}{\geq} \frac{2\left(1-t^i\right)}{r^{2i/K}} \prod_{k\in\mathcal{I}_i} p_k \\ &\geq \frac{2\left(1-t\right)}{r^{2i/K}} \prod_{k\in\mathcal{I}_i} p_k, \end{aligned} \tag{3.85}$$

where (k) follows using arguments similar to those made for (3.70).

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

$$\left\| \bigotimes_{k\in[K]} \mathbf{A}_k \right\|_F = \prod_{k\in[K]} \|\mathbf{A}_k\|_F = \left\| \bigotimes_{k\in[K]} \mathbf{A}_{\pi(k)} \right\|_F, \tag{3.86}$$

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

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

$$
\begin{aligned}
&= \Big\| (\mathbf{D}_{(1,l_1)} \otimes \cdots \otimes \mathbf{D}_{(K_1,l_{K_1})} \otimes \mathbf{D}_{(K_1+1,l_{K_1+1})} \otimes \cdots \otimes \mathbf{D}_{(K,l_K)}) \\
&\qquad - (\mathbf{D}_{(1,l'_1)} \otimes \cdots \otimes \mathbf{D}_{(K_1,l'_{K_1})} \otimes \mathbf{D}_{(K_1+1,l_{K_1+1})} \otimes \cdots \otimes \mathbf{D}_{(K,l_K)}) \Big\|_F^2 \\
&\overset{(l)}{=} \left\| \left(\bigotimes_{k\in[K_1]} \mathbf{D}_{(k,l_k)} - \bigotimes_{k\in[K_1]} \mathbf{D}_{(k,l'_k)}) \right) \otimes \mathbf{D}_{(K_1+1,l_{K_1+1})} \otimes \cdots \otimes \mathbf{D}_{(K,l_K)}) \right\|_F^2 \\
&= \left\| \bigotimes_{k\in[K_1]} \mathbf{D}_{(k,l_k)} - \bigotimes_{k\in[K_1]} \mathbf{D}_{(k,l'_k)} \right\|_F^2 \prod_{k=K_1+1}^{K} \left\| \mathbf{D}_{(k,l_k)} \right\|_F^2 \\
&= \left(\prod_{k=K_1+1}^{K} p_k \right) \left\| \sum_{\substack{\mathbf{i}\in\{0,1\}^{K_1} \\ \|\mathbf{i}\|_1\neq 0}} \eta^{K_1-\|\mathbf{i}\|_1} \nu^{\|\mathbf{i}\|_1} \left(\bigotimes_{k\in[K_1]} \mathbf{D}_{(k,i_k,l_k)} - \bigotimes_{k\in[K_1]} \mathbf{D}_{(k,i_k,l'_k)} \right) \right\|_F^2 \\
&\overset{(m)}{=} \left(\sum_{\substack{\mathbf{i}\in\{0,1\}^{K_1} \\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K_1-\|\mathbf{i}\|_1)} \nu^{2\|\mathbf{i}\|_1} \prod_{\substack{k\in[K_1] \\ i_k=0}} \|\mathbf{D}_k^*\|_F^2 \left\| \bigotimes_{\substack{k\in[K_1] \\ i_k=1}} \mathbf{D}_{(k,1,l_k)} - \bigotimes_{\substack{k\in[K_1] \\ i_k=1}} \mathbf{D}_{(k,1,l'_k)} \right\|_F^2 \right) \\
&\overset{(n)}{\geq} \left(\prod_{k=K_1+1}^{K} p_k \right) \left(\sum_{\substack{\mathbf{i}\in\{0,1\}^{K_1} \\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K_1-\|\mathbf{i}\|_1)} \nu^{2\|\mathbf{i}\|_1} \left(\prod_{\substack{k\in[K_1] \\ i_k=0}} p_k \right) \left(\frac{2}{r^{2\|\mathbf{i}\|_1/K}} \prod_{\substack{k\in[K_1] \\ i_k=1}} p_k \right) (1-t) \right) \\
&\quad \overset{(o)}{=} 2p\,(1-t) \sum_{k=0}^{K_1-1} \binom{K_1}{k} \left(1-\frac{\varepsilon'}{r^2}\right)^k \left(\frac{\varepsilon'}{r^2}\right)^{K_1-k} \\
&\quad \overset{(p)}{=} 2p\,(1-t) \left(1 - \left(1-\frac{\varepsilon'}{r^2}\right)^{K_1} \right) \\
&\quad \geq 2p\,(1-t) \left(1 - \left(1-\frac{\varepsilon'}{r^2}\right) \right) \\
&\quad = \frac{2p}{r^2}\,(1-t)\,\varepsilon', \qquad (3.87)
\end{aligned}
$$

where (l) follows from the distributive property of Kronecker products, (m) follows the fact that terms in the sum have orthogonal columns (from (1.6) and (3.68)), (n) follows from (3.85), (o) follows from substituting values for η and ν, and (p) follows from the binomial formula.

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

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

$$
\begin{aligned}
&\overset{(q)}{\leq} \sum_{\substack{\mathbf{i}\in\{0,1\}^K \\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)}\nu^{2\|\mathbf{i}\|_1}\left(\left\|\bigotimes_{k\in[K]}\mathbf{D}_{(k,i_k,l_k)}\right\|_F + \left\|\bigotimes_{k\in[K]}\mathbf{D}_{(k,i_k,l'_k)}\right\|_F\right)^2 \\
&= 4\sum_{\substack{\mathbf{i}\in\{0,1\}^K \\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)}\nu^{2\|\mathbf{i}\|_1}\left\|\bigotimes_{k\in[K]}\mathbf{D}_{(k,i_k,l_k)}\right\|_F^2 \\
&= 4\sum_{\substack{\mathbf{i}\in\{0,1\}^K \\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)}\nu^{2\|\mathbf{i}\|_1}\prod_{\substack{k\in[K] \\ i_k=0}}\|\mathbf{D}_k^*\|_F^2\prod_{\substack{k\in[K] \\ i_k=1}}\|\mathbf{D}_{(k,1,l_k)}\|_F^2 \\
&= 4\sum_{\substack{\mathbf{i}\in\{0,1\}^K \\ \|\mathbf{i}\|_1\neq 0}} \eta^{2(K-\|\mathbf{i}\|_1)}\nu^{2\|\mathbf{i}\|_1}\left(\prod_{\substack{k\in[K] \\ i_k=0}} p_k\right)\left(\prod_{\substack{k\in[K] \\ i_k=1}}\frac{p_k}{r^{2/K}}\right) \\
&\overset{(r)}{=} 4p\sum_{k=0}^{K-1}\binom{K}{k}\left(1-\frac{\varepsilon'}{r^2}\right)^k\left(\frac{\varepsilon'}{r^2}\right)^{K-k} \\
&\overset{(s)}{\leq} \frac{4Kp}{r^2}\varepsilon',
\end{aligned} \tag{3.88}
$$

where (q) follows from the triangle inequality, (r) follows from substituting values for η and ν, and (s) follows from similar arguments as in (3.83).

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

Assuming side information $\mathbf{T}(\mathbf{X})=\mathbf{X}$, conditioned on the coefficients $\mathbf{x}_n$, the observations $\mathbf{y}_n$ follow a multivariate Gaussian distribution with covariance matrix $\sigma^2\mathbf{I}$ and mean vector $\mathbf{D}\mathbf{x}_n$. From the convexity of the KL divergence [80], following similar arguments as in [1, 78], we have

$$
\begin{aligned}
I(\mathbf{Y};l|\mathbf{T}(\mathbf{X})) &= I(\mathbf{Y};l|\mathbf{X}) \\
&= \frac{1}{L}\sum_{l\in[L]}\mathbb{E}_{\mathbf{X}}\left\{D_{KL}\left(f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X})\middle\|\frac{1}{L}\sum_{l'\in[L]}f_{\mathbf{D}_{l'}}(\mathbf{Y}|\mathbf{X})\right)\right\} \\
&\leq \frac{1}{L^2}\sum_{l,l'\in[L]}\mathbb{E}_{\mathbf{X}}\left\{D_{KL}\left(f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X})\middle\|f_{\mathbf{D}_{l'}}(\mathbf{Y}|\mathbf{X})\right)\right\},
\end{aligned} \tag{3.89}
$$

where $f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X})$ is the probability distribution of the observations $\mathbf{Y}$, given the coefficient matrix $\mathbf{X}$ and the dictionary $\mathbf{D}_l$. From Durrieu et al. [81], we have

$$\begin{aligned} D_{KL}\Big(f_{\mathbf{D}_l}(\mathbf{Y}|\mathbf{X})\big\|f_{\mathbf{D}_{l'}}(\mathbf{Y}|\mathbf{X})\Big) &= \sum_{n\in[N]} \frac{1}{2\sigma^2}\,\|(\mathbf{D}_l-\mathbf{D}_{l'})\mathbf{x}_n\|_2^2 \\ &= \sum_{n\in[N]} \frac{1}{2\sigma^2}\,\mathrm{Tr}\left\{(\mathbf{D}_l-\mathbf{D}_{l'})^\top(\mathbf{D}_l-\mathbf{D}_{l'})\mathbf{x}_n\mathbf{x}_n^\top\right\}. \quad (3.90) \end{aligned}$$

Substituting (3.90) in (3.89) results in

$$\begin{aligned} I(\mathbf{Y};l|\mathbf{T}(\mathbf{X})) &\le \mathbb{E}_{\mathbf{X}}\left\{\sum_{n\in[N]} \frac{1}{2\sigma^2}\,\mathrm{Tr}\left\{(\mathbf{D}_l-\mathbf{D}_{l'})^\top(\mathbf{D}_l-\mathbf{D}_{l'})\mathbf{x}_n\mathbf{x}_n^\top\right\}\right\} \\ &= \sum_{n\in[N]} \frac{1}{2\sigma^2}\,\mathrm{Tr}\left\{(\mathbf{D}_l-\mathbf{D}_{l'})^\top(\mathbf{D}_l-\mathbf{D}_{l'})\mathbf{\Sigma}_x\right\} \\ &\overset{(t)}{\le} \sum_{n\in[N]} \frac{1}{2\sigma^2}\|\mathbf{\Sigma}_x\|_2\|\mathbf{D}_l-\mathbf{D}_{l'}\|_F^2 \\ &\overset{(u)}{\le} \frac{N}{2\sigma^2}\|\mathbf{\Sigma}_x\|_2\left(\frac{4Kp\varepsilon'}{r^2}\right) \\ &= \frac{2NKp\|\mathbf{\Sigma}_x\|_2}{r^2\sigma^2}\varepsilon', \quad (3.91) \end{aligned}$$

where (u) follows from (3.88). To show (t), we use the fact that for any $\mathbf{A}\in\mathbb{R}^{p\times p}$ and $\mathbf{\Sigma}_x$ with ordered singular values $\sigma_i(\mathbf{A})$ and $\sigma_i(\mathbf{\Sigma}_x), i\in[p]$, we have

$$\begin{aligned} \mathrm{Tr}\left\{\mathbf{A}\mathbf{\Sigma}_x\right\} &\le \left|\mathrm{Tr}\left\{\mathbf{A}\mathbf{\Sigma}_x\right\}\right| \\ &\overset{(v)}{\le} \sum_{i=1}^{p}\sigma_i(\mathbf{A})\sigma_i(\mathbf{\Sigma}_x) \\ &\overset{(w)}{\le} \sigma_1(\mathbf{\Sigma}_x)\sum_{i=1}^{p}\sigma_i(\mathbf{A}) \\ &= \|\mathbf{\Sigma}_x\|_2\,\mathrm{Tr}\{\mathbf{A}\}, \quad (3.92) \end{aligned}$$

where (v) follows from Von Neumann's trace inequality [82] and (w) follows from the positivity of the singular values of $\mathbf{\Sigma}_x$. The inequality in (t) follows from replacing $\mathbf{A}$ with $(\mathbf{D}_l-\mathbf{D}_{l'})^\top(\mathbf{D}_l-\mathbf{D}_{l'})$ and using the fact that $\mathrm{Tr}\{(\mathbf{D}_l-\mathbf{D}_{l'})^\top(\mathbf{D}_l-\mathbf{D}_{l'})\} = \|\mathbf{D}_l-\mathbf{D}_{l'}\|_F^2$.

3.8.3 Proof of Lemma 3.4

The dictionary class $\mathcal{D}_L$ constructed in Lemma 3.2 is again considered here. Note that (3.27) implies $\varepsilon' < r^2$, since $s \geq 1$. The first part of Lemma 3.4, up to (3.28), thus trivially follows from Lemma 3.2. In order to prove the second part, notice that in this case the coefficient vector is assumed to be sparse according to (3.10). Denoting $\mathbf{x}_{\mathcal{S}_n}$ as the elements of $\mathbf{x}_n$ with indices $\mathcal{S}_n \triangleq \text{supp}(\mathbf{x}_n)$, we have observations $\mathbf{y}_n$ as

$$\mathbf{y}_n = \mathbf{D}_{l,\mathcal{S}_n}\mathbf{x}_{\mathcal{S}_n} + \mathbf{w}_n. \tag{3.93}$$

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

$$\mathbf{\Sigma}_{(n,l)} = \sigma_a^2 \mathbf{D}_{l,\mathcal{S}_n}\mathbf{D}_{l,\mathcal{S}_n}^\top + \sigma^2 \mathbf{I}_s. \tag{3.94}$$

The conditional MI $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X}))$ has the following upper bound [1, 83]:

$$\begin{aligned} I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X})) &\leq \mathbb{E}_{\mathbf{T}(\mathbf{X})}\Bigg\{ \sum_{\substack{n\in[N]\\ l,l'\in[L]}} \frac{1}{L^2} \text{Tr}\left\{ \left[\mathbf{\Sigma}_{(n,l)}^{-1} - \mathbf{\Sigma}_{(n,l')}^{-1}\right]\left[\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\right]\right\}\Bigg\} \\ &\leq \text{rank}\left\{\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\right\} \mathbb{E}_{\mathbf{T}(\mathbf{X})}\Bigg\{ \sum_{n\in[N]} \frac{1}{L^2} \sum_{l,l'\in[L]} \left\|\mathbf{\Sigma}_{(n,l)}^{-1} - \mathbf{\Sigma}_{(n,l')}^{-1}\right\|_2 \left\|\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\right\|_2 \Bigg\}. \end{aligned} \tag{3.95}$$

Since $\text{rank}(\mathbf{\Sigma}_{(n,l)}) \leq s$, $\text{rank}\{\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\} \leq 2s$ [1].

Next, note that since non-zero elements of the coefficient vector are selected according to (3.10) and (3.25), we can write the subdictionary $\mathbf{D}_{l,\mathcal{S}_n}$ in terms of the Khatri-Rao product of matrices:

$$\mathbf{D}_{l,\mathcal{S}_n} = \bigast_{k\in[K]} \mathbf{D}_{(k,l_k),\mathcal{S}_{n_k}}, \tag{3.96}$$

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

multisets.[9] We can now write

$$\mathbf{\Sigma}_{(n,l)} = \sigma_a^2 \Bigg(\underset{k_1 \in [K]}{\circledast} \mathbf{D}_{(k_1,l_{k_1}),\mathcal{S}_{n_{k_1}}} \Bigg) \Bigg(\underset{k_2 \in [K]}{\circledast} \mathbf{D}_{(k_2,l_{k_2}),\mathcal{S}_{n_{k_2}}} \Bigg)^\top + \sigma^2 \mathbf{I}_s. \tag{3.97}$$

We next write

$$\begin{aligned}
\frac{1}{\sigma_a^2}(\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}) &= \Bigg(\underset{k_1 \in [K]}{\circledast} \mathbf{D}_{(k_1,l_{k_1}),\mathcal{S}_{n_{k_1}}} \Bigg) \Bigg(\underset{k_2 \in [K]}{\circledast} \mathbf{D}_{(k_2,l_{k_2}),\mathcal{S}_{n_{k_2}}} \Bigg)^\top \\
&\quad - \Bigg(\underset{k_1 \in [K]}{\circledast} \mathbf{D}_{(k_1,l'_{k_1}),\mathcal{S}_{n_{k_1}}} \Bigg) \Bigg(\underset{k_2 \in [K]}{\circledast} \mathbf{D}_{(k_2,l'_{k_2}),\mathcal{S}_{n_{k_2}}} \Bigg)^\top \\
&= \Bigg(\sum_{\mathbf{i} \in \{0,1\}^K} \eta^{K-\|\mathbf{i}\|_1} \nu^{\|\mathbf{i}\|_1} \underset{k_1 \in [K]}{\circledast} \mathbf{D}_{(k_1,i_{k_1},l_{k_1}),\mathcal{S}_{n_{k_1}}} \Bigg) \\
&\qquad \Bigg(\sum_{\mathbf{i}' \in \{0,1\}^K} \eta^{K-\|\mathbf{i}'\|_1} \nu^{\|\mathbf{i}'\|_1} \underset{k_2 \in [K]}{\circledast} \mathbf{D}_{(k_2,i'_{k_2},l_{k_2}),\mathcal{S}_{n_{k_2}}} \Bigg)^\top \\
&\quad - \Bigg(\sum_{\mathbf{i} \in \{0,1\}^K} \eta^{K-\|\mathbf{i}\|_1} \nu^{\|\mathbf{i}\|_1} \underset{k_1 \in [K]}{\circledast} \mathbf{D}_{(k_1,i_{k_1},l'_{k_1}),\mathcal{S}_{n_{k_1}}} \Bigg) \\
&\qquad \Bigg(\sum_{\mathbf{i}' \in \{0,1\}^K} \eta^{K-\|\mathbf{i}'\|_1} \nu^{\|\mathbf{i}'\|_1} \underset{k_2 \in [K]}{\circledast} \mathbf{D}_{(k_2,i'_{k_2},l'_{k_2}),\mathcal{S}_{n_{k_2}}} \Bigg)^\top \\
&= \sum_{\substack{\mathbf{i},\mathbf{i}' \in \{0,1\}^K \\ \|\mathbf{i}\|_1 + \|\mathbf{i}'\|_1 \neq 0}} \eta^{2K-\|\mathbf{i}\|_1-\|\mathbf{i}'\|_1} \nu^{\|\mathbf{i}\|_1+\|\mathbf{i}'\|_1} \\
&\quad \Bigg(\underset{k_1 \in [K]}{\circledast} \mathbf{D}_{(k_1,i_{k_1},l_{k_1}),\mathcal{S}_{n_{k_1}}} \Bigg) \Bigg(\underset{k_2 \in [K]}{\circledast} \mathbf{D}_{(k_2,i'_{k_2},l_{k_2}),\mathcal{S}_{n_{k_2}}} \Bigg)^\top \\
&\quad - \sum_{\substack{\mathbf{i},\mathbf{i}' \in \{0,1\}^K \\ \|\mathbf{i}\|_1 + \|\mathbf{i}'\|_1 \neq 0}} \eta^{2K-\|\mathbf{i}\|_1-\|\mathbf{i}'\|_1} \nu^{\|\mathbf{i}\|_1+\|\mathbf{i}'\|_1} \\
&\quad \Bigg(\underset{k_1 \in [K]}{\circledast} \mathbf{D}_{(k_1,i_{k_1},l'_{k_1}),\mathcal{S}_{n_{k_1}}} \Bigg) \Bigg(\underset{k_2 \in [K]}{\circledast} \mathbf{D}_{(k_2,i'_{k_2},l'_{k_2}),\mathcal{S}_{n_{k_2}}} \Bigg)^\top .
\end{aligned} \tag{3.98}$$

We now note that

$$\begin{aligned}
\|\mathbf{A}_1 * \mathbf{A}_2\|_2 &= \|(\mathbf{A}_1 \otimes \mathbf{A}_2)\mathbf{P}\|_2 \\
&\leq \|(\mathbf{A}_1 \otimes \mathbf{A}_2)\|_2 \|\mathbf{P}\|_2 \\
&\overset{(a)}{=} \|\mathbf{A}_1\|_2 \|\mathbf{A}_2\|_2,
\end{aligned} \tag{3.99}$$

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

where $\mathbf{P} \in \mathbb{R}^{p\times s}$ is a selection matrix that selects s columns of $\mathbf{A}_1 \otimes \mathbf{A}_2$ and $\mathbf{p}_j = \mathbf{e}_i$ for $j \in [s], i \in [p]$. Here, (a) follows from the fact that $\|\mathbf{P}\|_2 = 1$ ($\mathbf{P}^\top\mathbf{P} = \mathbf{I}_s$). From (3.27), it is apparent that $\sqrt{\frac{s\varepsilon'}{r^2}} \le 1$. Furthermore,

$$\left\|\mathbf{D}_{(k,0),\mathcal{S}_{n_k}}\right\|_2 \le \sqrt{\frac{3}{2}}, \left\|\mathbf{D}_{(k,1,l_k),\mathcal{S}_{n_k}}\right\|_2 \le \sqrt{\frac{s}{r^{2/K}}},\ k \in [K], \tag{3.100}$$

where the fist inequality in (3.100) follows from the RIP condition for $\left\{\mathbf{D}_{(0,k)}, k \in [K]\right\}$ and the second inequality follows from the fact that $\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_F$. We therefore have

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

$$
\begin{aligned}
&= 2\sqrt{\frac{s\varepsilon'}{r^2}}\Big(\sum_{k_1=0}^{K-1}\binom{K}{k_1}\Big(\sqrt{\frac{3}{2}}\Big)^{k_1}\Big(\sqrt{\frac{s\varepsilon'}{r^2}}\Big)^{K-1-k_1}\Big)\\
&\quad\Big(\sum_{k_2=0}^{K}\binom{K}{k_2}\Big(\sqrt{\frac{3}{2}}\Big)^{k_2}\Big(\sqrt{\frac{s\varepsilon'}{r^2}}\Big)^{K-k_2}+\Big(\sqrt{\frac{3}{2}}\Big)^{K}\Big)\\
&\overset{(f)}{\leq} 2\sqrt{\frac{s\varepsilon'}{r^2}}\Big(\Big(\sqrt{\frac{3}{2}}\Big)^{K-1}\sum_{k_1=0}^{K}\binom{K}{k_1}\Big)\Big(\Big(\sqrt{\frac{3}{2}}+1\Big)^{K}+\Big(\sqrt{\frac{3}{2}}\Big)^{K}\Big)\\
&\leq 2\sqrt{\frac{s\varepsilon'}{r^2}}\Big(\Big(\sqrt{\frac{3}{2}}\Big)^{K-1}2^K\Big)\Big(\Big(\frac{3}{2}\Big)^{K}2^K+\Big(\frac{3}{2}\Big)^{K}\Big)\\
&\leq 3^{2K+1}\sqrt{\frac{s\varepsilon'}{r^2}},
\end{aligned}
\tag{3.101}
$$

where (b) follows from triangle inequality, (c) follows from (3.99), (d) follows from (3.100), (e) and (f) follow from replacing the value for ν and the fact that $\eta < 1$ and $s\varepsilon'/r^2 < 1$ (by assumption). Denoting the smallest eigenvalue of $\mathbf{\Sigma}_{(n,l)}$ as $\lambda_{\min}(\mathbf{\Sigma}_{(n,l)})$, $\lambda_{\min}(\mathbf{\Sigma}_{(n,l)}) \geq \sigma^2$ holds; thus, we have $\|\mathbf{\Sigma}_{(n,l)}^{-1}\|_2 \leq \frac{1}{\sigma^2}$ and from [84], we get

$$
\begin{aligned}
\left\|\mathbf{\Sigma}_{(n,l)}^{-1}-\mathbf{\Sigma}_{(n,l')}^{-1}\right\|_2 &\leq 2\left\|\mathbf{\Sigma}_{(n,l)}^{-1}\right\|_2^2\left\|\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}\right\|_2\\
&\leq \frac{2}{\sigma^4}\left\|\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}\right\|_2.
\end{aligned}
\tag{3.102}
$$

Now (3.95) can be stated as

$$
\begin{aligned}
I(\mathbf{Y};l|\mathbf{T}(\mathbf{X})) &\leq \frac{4Ns}{\sigma^4L^2}\sum_{l,l'}\left\|\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}\right\|_2^2\\
&\leq \frac{4Ns}{\sigma^4}\left\|\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}\right\|_2^2\\
&\overset{(g)}{\leq} \frac{4Ns}{\sigma^4}(3^{4K+2})\left(\sigma_a^2\sqrt{\frac{s\varepsilon'}{r^2}}\right)^2\\
&= 36(3^{4K})\left(\frac{\sigma_a}{\sigma}\right)^4\frac{Ns^2}{r^2}\varepsilon',
\end{aligned}
\tag{3.103}
$$

where (g) follow from (3.101). Thus, the proof is complete.

3.8.4 Proof of Lemma 3.5

Similar to Lemma 3.4, the first part of this Lemma trivially follows from Lemma 3.2. Also, in this case the coefficient vector is assumed to be sparse according to (3.11). Hence, conditioned on $\mathcal{S}_n = \text{supp}(\mathbf{x}_n)$, observations $\mathbf{y}_n$'s are zero-mean independent multivariate Gaussian random vectors with covariances given by (3.94). Similar to Lemma 3.4, therefore, the conditional MI has the upper bound given in (3.95). We now simplify this upper bound further.

When non-zero elements of the coefficient vector are selected according to (3.11) and (3.25), we can write the dictionary $\mathbf{D}_{l,\mathcal{S}_n}$ in terms of the Kronecker product of matrices:

$$\mathbf{D}_{l,\mathcal{S}_n} = \bigotimes_{k\in[K]} \mathbf{D}_{(k,l_k),\mathcal{S}_{n_k}}, \tag{3.104}$$

where $\mathcal{S}_{n_k} = \{j_{n_k}\}_{n_k=1}^{s_k}, j_{n_k} \in [p_k]$, for all $k \in [K]$, denotes the support of $\mathbf{x}_n$ on coordinate dictionary $\mathbf{D}_{(k,l_k)}$ and $\mathcal{S}_n$ corresponds to indexing of the elements of $(\mathcal{S}_1 \times \cdots \times \mathcal{S}_K)$. Note that $\mathbf{D}_{l,\mathcal{S}_n} \in \mathbb{R}^{(\prod_{k\in[K]} m_k)\times s}$. In contrast to coefficient model (3.10), in this model the $\mathcal{S}_{n_k}$'s are not multisets anymore since for each $\mathbf{D}_{(k,l_k)}, k \in [K]$, we select s_k columns at random and $\mathbf{D}_{(k,l_k),\mathcal{S}_{n_k}}$ are submatrices of $\mathbf{D}_{(k,l_k)}$. Therefore, (3.94) can be written as

$$\mathbf{\Sigma}_{(n,l)} = \sigma_a^2 \Bigg(\bigotimes_{k_1\in[K]} \mathbf{D}_{(k_1,l_{k_1}),\mathcal{S}_{n_{k_1}}} \Bigg) \Bigg(\bigotimes_{k_2\in[K]} \mathbf{D}_{(k_2,l_{k_2}),\mathcal{S}_{n_{k_2}}} \Bigg)^\top + \sigma^2 \mathbf{I}_s. \tag{3.105}$$

In order to find an upper bound for $\|\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}\|_2$, notice that the expression for $\mathbf{\Sigma}_{(n,l)} - \mathbf{\Sigma}_{(n,l')}$ is similar to that of (3.98), where $\divideontimes$ is replaced by $\bigotimes$. Using the property of Kronecker product that $\|\mathbf{A}_1 \otimes \mathbf{A}_2\|_2 = \|\mathbf{A}_1\|_2 \|\mathbf{A}_2\|_2$ and the fact that

$$\left\|\mathbf{D}_{(k,0),\mathcal{S}_{n_k}}\right\|_2 \leq \sqrt{\frac{3}{2}}, \left\|\mathbf{D}_{(k,1,l_k),\mathcal{S}_{n_k}}\right\|_2 \leq \sqrt{\frac{s_k}{r^{2/K}}}, \forall k \in [K], \tag{3.106}$$

we have

$$
\begin{aligned}
&\frac{1}{\sigma_a^2}\left\|\mathbf{\Sigma}_{(n,l)}-\mathbf{\Sigma}_{(n,l')}\right\|_2 \\
&\leq 2\sum_{\substack{\mathbf{i},\mathbf{i}'\in\{0,1\}^K\\ \|\mathbf{i}\|_1+\|\mathbf{i}'\|_1\neq 0}} \eta^{2K-\|\mathbf{i}\|_1-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}\|_1+\|\mathbf{i}'\|_1}\left\|\bigotimes_{k_1\in[K]}\mathbf{D}_{(k_1,i_{k_1},l_{k_1}),\mathcal{S}_{n_{k_1}}}\right\|_2\left\|\bigotimes_{k_2\in[K]}\mathbf{D}_{(k_2,i'_{k_2},l_{k_2}),\mathcal{S}_{n_{k_2}}}\right\|_2 \\
&= 2\sum_{\substack{\mathbf{i}\in\{0,1\}^K\\ \|\mathbf{i}\|_1\neq 0}} \eta^{K-\|\mathbf{i}\|_1}\nu^{\|\mathbf{i}\|_1}\prod_{\substack{k_1\in[K]\\ i_{k_1}=0}}\left\|\mathbf{D}_{(k_1,0),\mathcal{S}_{n_{k_1}}}\right\|_2\prod_{\substack{k_1\in[K]\\ i_{k_1}=1}}\left\|\mathbf{D}_{(k_1,1,l_{k_1}),\mathcal{S}_{n_{k_1}}}\right\|_2 \\
&\quad\left(\sum_{\mathbf{i}'\in\{0,1\}^K}\eta^{K-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}'\|_1}\prod_{\substack{k_2\in[K]\\ i'_{k_2}=0}}\left\|\mathbf{D}_{(k_2,0),\mathcal{S}_{n_{k_2}}}\right\|_2\prod_{\substack{k_2\in[K]\\ i'_{k_2}=1}}\left\|\mathbf{D}_{(k_2,1,l_{k_2}),\mathcal{S}_{n_{k_2}}}\right\|_2\right) \\
&+2\left(\eta^K\prod_{k_1\in[K]}\left\|\mathbf{D}_{(k_1,0),\mathcal{S}_{n_{k_1}}}\right\|_2\right)\left(\sum_{\substack{\mathbf{i}'\in\{0,1\}^K\\ \|\mathbf{i}'\|_1\neq 0}}\eta^{K-\|\mathbf{i}'\|_1}\nu^{\|\mathbf{i}'\|_1}\right. \\
&\qquad\left.\prod_{\substack{k_2\in[K]\\ i'_{k_2}=0}}\left\|\mathbf{D}_{(k_2,0),\mathcal{S}_{n_{k_2}}}\right\|_2\prod_{\substack{k_2\in[K]\\ i'_{k_2}=1}}\left\|\mathbf{D}_{(k_2,1,l_{k_2}),\mathcal{S}_{n_{k_2}}}\right\|_2\right) \\
&\overset{(a)}{\leq} 2\sqrt{s}\Bigg[\left(\sum_{k_1=0}^{K-1}\binom{K}{k_1}\eta^{k_1}\nu^{K-k_1}\left(\sqrt{\frac{3}{2}}\right)^{k_1}\left(\sqrt{\frac{1}{r^{2/K}}}\right)^{K-k_1}\right) \\
&\quad\left(\sum_{k_2=0}^{K}\binom{K}{k_2}\left(\eta\sqrt{\frac{3}{2}}\right)^{k_2}\right)+\left(\eta\sqrt{\frac{3}{2}}\right)^K \\
&\quad\left(\sum_{k_2=0}^{K-1}\binom{K}{k_2}\eta^{k_2}\nu^{(K-k_2)}\left(\sqrt{\frac{3}{2}}\right)^{k_2}\left(\sqrt{\frac{1}{r^{2/K}}}\right)^{K-k_2}\right)\Bigg] \\
&\overset{(b)}{\leq} 2\sqrt{\frac{s\varepsilon'}{r^2}}\left(\sum_{k_1=0}^{K-1}\binom{K}{k_1}\left(\sqrt{\frac{3}{2}}\right)^{k_1}\right)\left(\left(\sum_{k_2=0}^{K}\binom{K}{k_2}\left(\sqrt{\frac{3}{2}}\right)^{k_2}\right)+\left(\sqrt{\frac{3}{2}}\right)^K\right) \\
&\overset{(c)}{\leq} 3^{2K+1}\sqrt{\frac{s\varepsilon'}{r^2}},
\end{aligned}
\tag{3.107}
$$

where (a) follows from (3.106), (b) follows from replacing the value for ν and the fact that $\eta < 1$, $\varepsilon'/r^2 < 1$ (by assumption), and (c) follows from similar arguments in (3.101). The rest of the proof follows the same arguments as in Lemma 3.4 and (3.103) holds in this case as well.

3.8.5 Proof of Theorem 3.4

Any dictionary $\mathbf{D}^0 \in \mathcal{X}(\mathbf{I}_p, r)$ can be written as

$$\begin{aligned}\mathbf{D}^0 &= \mathbf{A} \otimes \mathbf{B} \\ &= (\mathbf{I}_{p_1} + \mathbf{\Delta}_1) \otimes (\mathbf{I}_{p_2} + \mathbf{\Delta}_2),\end{aligned} \tag{3.108}$$

We have to ensure that $\|\mathbf{D}^0 - \mathbf{I}_p\|_F \leq r$. We have

$$\begin{aligned}\|\mathbf{D}^0 - \mathbf{I}_p\|_F &= \|\mathbf{I}_{p_1} \otimes \mathbf{\Delta}_2 + \mathbf{\Delta}_1 \otimes \mathbf{I}_{p_2} + \mathbf{\Delta}_1 \otimes \mathbf{\Delta}_2\|_F \\ &\leq \|\mathbf{I}_{p_1} \otimes \mathbf{\Delta}_2\|_F + \|\mathbf{\Delta}_1 \otimes \mathbf{I}_{p_2}\|_F + \|\mathbf{\Delta}_1 \otimes \mathbf{\Delta}_2\|_F \\ &= \|\mathbf{I}_{p_1}\|_F \|\mathbf{\Delta}_2\|_F + \|\mathbf{\Delta}_1\|_F \|\mathbf{I}_{p_2}\|_F + \|\mathbf{\Delta}_1\|_F \|\mathbf{\Delta}_2\|_F \\ &\leq r_2\sqrt{p_1} + r_1\sqrt{p_2} + r_1 r_2 \\ &\overset{(a)}{\leq} r,\end{aligned} \tag{3.109}$$

where (a) follows from (3.37). Therefore, we have

$$\begin{aligned}\mathbf{D}^0 \in \Big\{ &\mathbf{A} \otimes \mathbf{B} = (\mathbf{I}_{p_1} + \mathbf{\Delta}_1) \otimes (\mathbf{I}_{p_2} + \mathbf{\Delta}_2) \big|\ \|\mathbf{\Delta}_1\|_F \leq r_1, \|\mathbf{\Delta}_2\|_F \leq r_2, \\ &r_2\sqrt{p_1} + r_1\sqrt{p_2} + r_1 r_2 \leq r, \|\mathbf{a}_{l_1}\|_2 = 1, l_1 \in [p_1],\ \|\mathbf{b}_{l_2}\|_2 = 1, l_2 \in [p_2] \Big\}.\end{aligned} \tag{3.110}$$

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

$$\mathbf{y}'_{(n,j)} = \mathbf{A}\mathbf{x}'_{(n,j)} + \mathbf{A}_p \mathbf{x}_n,\ j \in [p_2],\ n \in [N], \tag{3.111}$$

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

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

$$\mathbf{y}''_{(n,j)} = \mathbf{B}\mathbf{x}''_{(n,j)} + \mathbf{B}_p \mathbf{x}_n,\ j \in [p_1],\ n \in [N], \tag{3.112}$$

where $\mathbf{B}_p \triangleq (\mathbf{\Delta}_1 \otimes \mathbf{B})^{\mathcal{I}_n}$ denotes the matrix consisting of the rows of $(\mathbf{\Delta}_1 \otimes \mathbf{B})$ with

indices $\mathcal{I}_n \triangleq jp_2 + i$, where $i = \{0\} \cup [p_2 - 1]$ and $j = (n-1) \bmod p_1$. Given the fact that $\mathbf{x}_n \in \{-1, 0, 1\}^p$, $\sigma_a^2 = 1$ and $\|\mathbf{x}_n\|_2^2 = s$, after division of the coefficient vector according to (3.41) and (3.45), we have

$$\mathbb{E}_{\mathbf{x}_n}\left\{x_{n,l}^2\right\} = \mathbb{E}_{\mathbf{x}'_{(n,j_1)}}\left\{x'^2_{(n,j_1),l_1}\right\} = \mathbb{E}_{\mathbf{x}''_{(n,j_2)}}\left\{x''^2_{(n,j_2),l_2}\right\} = \frac{s}{p}, \tag{3.113}$$

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

$$\text{SNR} = \frac{\mathbb{E}_{\mathbf{x}}\left\{\|\mathbf{x}\|_2^2\right\}}{\mathbb{E}_{\mathbf{w}}\left\{\|\mathbf{w}\|_2^2\right\}} = \frac{s}{m\sigma^2}. \tag{3.114}$$

We are interested in upper bounding $\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}^0\right\|_F^2\right\}$. For this purpose we first upper bound $\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{A}}(\mathbf{Y}) - \mathbf{A}\right\|_F^2\right\}$ and $\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{B}}(\mathbf{Y}) - \mathbf{B}\right\|_F^2\right\}$. We can split these MSEs into the sum of column-wise MSEs:

$$\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{A}}(\mathbf{Y}) - \mathbf{A}\right\|_F^2\right\} = \sum_{l=1}^{p_1} \mathbb{E}_{\mathbf{Y}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y}) - \mathbf{a}_l\|_2^2\right\}. \tag{3.115}$$

By construction:

$$\begin{aligned}\|\widehat{\mathbf{a}}_l(\mathbf{Y}) - \mathbf{a}_l\|_2^2 &\leq 2\left(\|\widehat{\mathbf{a}}_l(\mathbf{Y})\|_2^2 + \|\mathbf{a}_l\|_2^2\right) \\ &\overset{(b)}{\leq} 4,\end{aligned} \tag{3.116}$$

where (b) follows from the projection step in (3.48). We define the event $\mathcal{C}$ to be

$$\mathcal{C} \triangleq \bigcap_{\substack{n\in[N]\\ l\in[p]}} \{|w_{n,l}| \leq 0.4\}. \tag{3.117}$$

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

$$y_{n,l} - x_{n,l} = \left(\mathbf{I}_{p_1} \otimes \boldsymbol{\Delta}_2 + \boldsymbol{\Delta}_1 \otimes \mathbf{I}_{p_2} + \boldsymbol{\Delta}_1 \otimes \boldsymbol{\Delta}_2\right)^l \mathbf{x}_n + w_{n,l} \tag{3.118}$$

and

$$\begin{aligned}
&\left|\left(\mathbf{I}_{p_1}\otimes\boldsymbol{\Delta}_2+\boldsymbol{\Delta}_1\otimes\mathbf{I}_{p_2}+\boldsymbol{\Delta}_1\otimes\boldsymbol{\Delta}_2\right)^l\mathbf{x}_n+w_{n,l}\right| \\
&\quad\le\left\|\left(\mathbf{I}_{p_1}\otimes\boldsymbol{\Delta}_2+\boldsymbol{\Delta}_1\otimes\mathbf{I}_{p_2}+\boldsymbol{\Delta}_1\otimes\boldsymbol{\Delta}_2\right)^l\right\|_2\|\mathbf{x}_n\|_2+|w_{n,l}| \\
&\quad\le\left(\|\boldsymbol{\Delta}_1\|_F+\|\boldsymbol{\Delta}_2\|_F+\|\boldsymbol{\Delta}_1\|_F\|\boldsymbol{\Delta}_2\|_F\right)\|\mathbf{x}_n\|_2+|w_{n,l}| \\
&\quad\le(r_1+r_2+r_1r_2)\sqrt{s}+|w_{n,l}|. \qquad (3.119)
\end{aligned}$$

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

$$\begin{cases} y_{n,l}>0.5 & \text{if } x_{n,l}=1, \\ -0.5<y_{n,l}<0.5 & \text{if } x_{n,l}=0, \\ y_{n,l}<-0.5 & \text{if } x_{n,l}=-1, \end{cases} \qquad (3.120)$$

thus, ensuring correct recovery of coefficients ($\widehat{\mathbf{X}}=\mathbf{X}$) using the thresholding technique (3.39) when conditioned on $\mathcal{C}$. Using standard tail bounds for Gaussian random variables [1, (92)], [85, Proposition 7.5] and taking a union bound over all pN i.i.d. variables $\{w_{n,l}\}, n\in[N], l\in[p]$, we have

$$\mathbb{P}\left\{\mathcal{C}^c\right\}\le\exp\left(-\frac{0.08pN}{\sigma^2}\right). \qquad (3.121)$$

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

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Y}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\right\} \\
&\quad=\mathbb{E}_{\mathbf{Y},\mathbf{W}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\,|\mathcal{C}\right\}\mathbb{P}(\mathcal{C})+\mathbb{E}_{\mathbf{Y},\mathbf{W}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\,|\mathcal{C}^c\right\}\mathbb{P}(\mathcal{C}^c) \\
&\quad\overset{(c)}{\le}\mathbb{E}_{\mathbf{Y},\mathbf{W}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\,|\mathcal{C}\right\}+4\exp\left(-\frac{0.08pN}{\sigma^2}\right), \qquad (3.122)
\end{aligned}$$

where (c) follows from (3.116) and (3.121). To bound $\mathbb{E}_{\mathbf{Y},\mathbf{W}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\,|\mathcal{C}\right\}$, we

have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{Y},\mathbf{W}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\,|\mathcal{C}\right\}\\
&\quad=\mathbb{E}_{\mathbf{Y},\mathbf{W}}\left\{\|P_{\mathcal{B}_1}(\widetilde{\mathbf{a}}_l(\mathbf{Y}))-\mathbf{a}_l\|_2^2\,|\mathcal{C}\right\}\\
&\quad\overset{(d)}{\leq}\mathbb{E}_{\mathbf{Y},\mathbf{W}}\left\{\|\widetilde{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\,|\mathcal{C}\right\}\\
&\quad\overset{(e)}{=}\mathbb{E}_{\mathbf{Y},\mathbf{W}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}\widehat{x'}_{(n,j),l}\mathbf{y}'_{(n,j)}-\mathbf{a}_l\right\|_2^2\Big|\mathcal{C}\right\}\\
&\quad\overset{(f)}{=}\mathbb{E}_{\mathbf{Y},\mathbf{X},\mathbf{W}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\mathbf{y}'_{(n,j)}-\mathbf{a}_l\right\|_2^2\Big|\mathcal{C}\right\}\\
&\quad\overset{(g)}{=}\mathbb{E}_{\mathbf{X},\mathbf{W}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\big(\mathbf{A}\mathbf{x}'_{(n,j)}+\mathbf{A}_p\mathbf{x}_n+\mathbf{w}'_{(n,j)}\big)-\mathbf{a}_l\right\|_2^2\Big|\mathcal{C}\right\}\\
&\quad\overset{(h)}{\leq}2\mathbb{E}_{\mathbf{X},\mathbf{W}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\mathbf{w}'_{(n,j)}\right\|_2^2\Big|\mathcal{C}\right\}\\
&\quad+4\mathbb{E}_{\mathbf{X},\mathbf{W}}\left\{\left\|\mathbf{a}_l-\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\sum_{t=1}^{p_1}\mathbf{a}_t x'_{(n,j),t}\right\|_2^2\Big|\mathcal{C}\right\}\\
&\quad+4\mathbb{E}_{\mathbf{X},\mathbf{W}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\sum_{t=1}^{p}\mathbf{a}_{p,t}x_{n,t}\right\|_2^2\Big|\mathcal{C}\right\},
\end{aligned}
\tag{3.123}
$$

where (d) follows from the fact that $\|\mathbf{a}_l\|_2=1$, (e) follows from (3.43), (f) follows from the fact that conditioned on the event $\mathcal{C}$, $\widehat{\mathbf{X}}=\mathbf{X}$, (g) follows from (3.111) and (h) follows from the fact that $\|\mathbf{x}_1+\mathbf{x}_2\|_2^2\leq 2(\|\mathbf{x}_1\|_2^2+\|\mathbf{x}_2\|_2^2)$. We bound the three terms in (3.123) separately. Defining $\nu\triangleq\mathcal{Q}(-0.4/\sigma)-\mathcal{Q}(0.4/\sigma)$, where $\mathcal{Q}(x)\triangleq\int_{z=x}^{\infty}\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})dz$, we can bound the noise variance conditioned on $\mathcal{C}$, $\sigma^2_{w_{n,t}}$, by [1]

$$
\sigma^2_{w_{n,t}}\leq\frac{\sigma^2}{\nu}. \tag{3.124}
$$

The first expectation in (3.123) can be bounded by

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{X},\mathbf{W}}\left\{\left\|\frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\mathbf{w}'_{(n,j)}\right\|_2^2\Big|\mathcal{C}\right\}\\
&\quad=\left(\frac{p_1}{Ns}\right)^2\sum_{n,n'=1}^{N}\sum_{j,j'=1}^{p_2}\mathbb{E}_{\mathbf{X},\mathbf{W}}\left\{x'_{(n,j),l}x'_{(n',j'),l}w'^{\top}_{(n',j')}w'_{(n,j)}|\mathcal{C}\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\frac{p_1}{Ns}\right)^2 \sum_{n=1}^{N}\sum_{j=1}^{p_2}\sum_{t=1}^{m_1} \mathbb{E}_{\mathbf{X},\mathbf{W}}\left\{x'^2_{(n,j),l}|\mathcal{C}\right\}\mathbb{E}_{\mathbf{X},\mathbf{W}}\left\{w'^2_{(n,j),t}|\mathcal{C}\right\} \\
&\overset{(i)}{=} \left(\frac{p_1}{Ns}\right)^2 Np_2\mathbb{E}_{\mathbf{X}}\left\{x'^2_{(n,j),l}\right\}\mathbb{E}_{\mathbf{W}}\left\{w'^2_{(n,j),t}|\mathcal{C}\right\} \\
&\overset{(j)}{\leq} \left(\frac{p_1}{Ns}\right)^2 Np_2\left(\frac{s}{p}\right)\left(\frac{m_1\sigma^2}{\nu}\right) \\
&\overset{(k)}{\leq} \frac{2m_1p_1\sigma^2}{Ns},
\end{aligned} \tag{3.125}
$$

where (i) follows from the fact that $\mathbf{x}'_{(n,j)}$ is independent of the event $\mathcal{C}$, (j) follows from (3.113) and (3.124), and (k) follows from the fact that $\nu \geq 0.5$ under the assumption that $\sigma \leq 0.4$ [1].

To bound the second expectation in (3.123), we use similar arguments as in Jung et al. [1]. We can write

$$
\mathbb{E}_{\mathbf{X}}\{x'_{(n,j),l}x'_{(n,j),t}x'_{(n',j'),l}x'_{(n',j'),t'}\} = \begin{cases} (\frac{s}{p})^2 & \text{if } (n,j)=(n',j') \text{ and } t=t'\neq l, \\ (\frac{s}{p})^2 & \text{if } (n,j)\neq(n',j') \text{ and } t=t'=l, \\ \frac{s}{p} & \text{if } (n,j)=(n',j') \text{ and } t=t'=l, \\ 0 & \text{otherwise,} \end{cases} \tag{3.126}
$$

and we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{X},\mathbf{W}}\left\{\left\|\mathbf{a}_l - \frac{p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}x'_{(n,j),l}\sum_{t=1}^{p_1}\mathbf{a}_t x'_{(n,j),t}\right\|_2^2\middle|\mathcal{C}\right\} \\
&\quad\leq \mathbf{a}_l^\top\mathbf{a}_l - \frac{2p_1}{Ns}\sum_{n=1}^{N}\sum_{j=1}^{p_2}\sum_{t=1}^{p_1}\mathbf{a}_l^\top\mathbf{a}_t\mathbb{E}_{\mathbf{X}}\left\{x'_{(n,j),l}x'_{(n,j),t}\right\} \\
&\qquad + \left(\frac{p_1}{Ns}\right)^2\sum_{n,n'=1}^{N}\sum_{j,j'=1}^{p_2}\sum_{t,t'=1}^{p_1}\mathbf{a}_{t'}^\top\mathbf{a}_t\mathbb{E}_{\mathbf{X}}\left\{x'_{(n',j'),l}x'_{(n',j'),t'}x'_{(n,j),l}x'_{(n,j),t}\right\} \\
&\quad= 1 - \left(\frac{2p_1}{Ns}\right)(p_2N)\left(\frac{s}{p}\right) + \left(\frac{p_1}{Ns}\right)^2(p_2N)\left(\frac{s}{p} + (p_1-1)\left(\frac{s}{p}\right)^2 + (p_2N-1)\left(\frac{s}{p}\right)^2\right) \\
&\quad= \frac{p_1}{N}\left(\frac{1}{s}+\frac{1}{p_2}-\frac{2}{p}\right) \\
&\quad\leq \frac{2p_1}{N}.
\end{aligned} \tag{3.127}
$$

To upper bound the third expectation in (3.123), we need to bound the ℓ_2 norm of columns of $\mathbf{A}_p$. We have

$$\begin{aligned}\forall t \in [p] : \|\mathbf{a}_{p,t}\|_2^2 &\overset{(l)}{\leq} \|(\mathbf{A} \otimes \mathbf{\Delta}_2)_t\|_2^2 \\ &\leq \|\mathbf{a}_l\|_2^2 \|\mathbf{\Delta}_2\|_F^2 \\ &= r_2^2, \end{aligned} \tag{3.128}$$

where $(\mathbf{A} \otimes \mathbf{\Delta}_2)_t$ denotes the t-th column of $(\mathbf{A} \otimes \mathbf{\Delta}_2)$ and (l) follows from the fact that $\mathbf{A}_p$ is a submatrix of $(\mathbf{A} \otimes \mathbf{\Delta}_2)$. Moreover, similar to the expectation in (3.126), we have

$$\mathbb{E}_{\mathbf{X}}\left\{x'_{(n,j),l} x'_{(n',j'),l} x_{n,t} x_{n',t'}\right\} = \begin{cases} (\frac{s}{p})^2 & \text{if } (n,j) = (n',j') \text{ and } t = t' \neq l', \\ (\frac{s}{p})^2 & \text{if } (n,j) \neq (n',j') \text{ and } t = t' = l', \\ \frac{s}{p} & \text{if } (n,j) = (n',j') \text{ and } t = t' = l', \\ 0 & \text{Otherwise,} \end{cases} \tag{3.129}$$

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

$$\begin{aligned} &\mathbb{E}_{\mathbf{X},\mathbf{W}}\left\{\left\| \frac{p_1}{Ns} \sum_{n=1}^{N} \sum_{j=1}^{p_2} x'_{(n,j),l} \sum_{t=1}^{p} \mathbf{a}_{p,t} x_{n,t} \right\|_2^2 \middle| \mathcal{C} \right\} \\ &= \left(\frac{p_1}{Ns}\right)^2 \sum_{n,n'=1}^{N} \sum_{j,j'=1}^{p_2} \sum_{t,t'=1}^{p} \mathbf{a}_{p,t'}^\top \mathbf{a}_{p,t} \mathbb{E}_{\mathbf{X}}\left\{x'_{(n,j),l} x'_{(n',j'),l} x_{n,t} x_{n',t'}\right\} \\ &\overset{(m)}{\leq} r_2^2 \left(\frac{p_1}{Ns}\right)^2 N p_2 \left(\frac{s}{p} + (p-1)\left(\frac{s}{p}\right)^2 + (Np_2 - 1)\left(\frac{s}{p}\right)^2\right) \\ &\leq r_2^2 \left(\frac{p_1}{Ns} + \frac{p_1}{N} + 1\right) \\ &\overset{(n)}{\leq} \frac{p_1}{N}, \end{aligned} \tag{3.130}$$

where (m) follows from (3.128) and (n) follows from the assumption in (3.37). Summing

up (3.125), (3.127), and (3.130), we have

$$\mathbb{E}_{\mathbf{Y}}\left\{\|\widehat{\mathbf{a}}_l(\mathbf{Y})-\mathbf{a}_l\|_2^2\right\} \leq \frac{4p_1}{N}\left(\frac{m_1\sigma^2}{s}+3\right)+4\exp\left(-\frac{0.08pN}{\sigma^2}\right). \tag{3.131}$$

Summing up the MSE for all columns, we obtain:

$$\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{A}}(\mathbf{Y})-\mathbf{A}\right\|_F^2\right\} \leq \frac{4p_1^2}{N}\left(\frac{m_1\sigma^2}{s}+3\right)+4p_1\exp\left(-\frac{0.08pN}{\sigma^2}\right). \tag{3.132}$$

We can follow similar steps to get

$$\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{B}}(\mathbf{Y})-\mathbf{B}\right\|_F^2\right\} \leq \frac{4p_2^2}{N}\left(\frac{m_2\sigma^2}{s}+3\right)+4p_2\exp\left(-\frac{0.08pN}{\sigma^2}\right). \tag{3.133}$$

From (3.132) and (3.133), we get

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{D}}(\mathbf{Y})-\mathbf{D}^0\right\|_F^2\right\}\\
&\quad=\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{A}}(\mathbf{Y})\otimes\widehat{\mathbf{B}}(\mathbf{Y})-\mathbf{A}\otimes\mathbf{B}\right\|_F^2\right\}\\
&\quad=\mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{A}}(\mathbf{Y})-\mathbf{A})\otimes\widehat{\mathbf{B}}(\mathbf{Y})+\mathbf{A}\otimes(\widehat{\mathbf{B}}(\mathbf{Y})-\mathbf{B})\right\|_F^2\right\}\\
&\quad\leq 2\left(\mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{A}}(\mathbf{Y})-\mathbf{A})\otimes\widehat{\mathbf{B}}(\mathbf{Y})\right\|_F^2\right\}+\mathbb{E}_{\mathbf{Y}}\left\{\left\|\mathbf{A}\otimes(\widehat{\mathbf{B}}(\mathbf{Y})-\mathbf{B})\right\|_F^2\right\}\right)\\
&\quad\leq 2\left(\mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{A}}(\mathbf{Y})-\mathbf{A})\right\|_F^2\right\}\mathbb{E}_{\mathbf{Y}}\left\{\left\|\widehat{\mathbf{B}}(\mathbf{Y})\right\|_F^2\right\}+\|\mathbf{A}\|_F^2\,\mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{B}}(\mathbf{Y})-\mathbf{B})\right\|_F^2\right\}\right)\\
&\quad\leq 2\left(p_2\mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{A}}(\mathbf{Y})-\mathbf{A})\right\|_F^2\right\}+p_1\mathbb{E}_{\mathbf{Y}}\left\{\left\|(\widehat{\mathbf{B}}(\mathbf{Y})-\mathbf{B})\right\|_F^2\right\}\right)\\
&\quad\leq \frac{8p}{N}\left(\frac{\sigma^2}{s}\sum_{k=1}^{2}m_kp_k+3\sum_{k=1}^{2}p_k\right)+8p\exp\left(-\frac{0.08pN}{\sigma^2}\right)\\
&\quad\overset{(o)}{=} \frac{8p}{N}\left(\frac{\sum_{k=1}^{2}m_kp_k}{m\,\mathrm{SNR}}+3\sum_{k=1}^{2}p_k\right)+8p\exp\left(-\frac{0.08pN}{\sigma^2}\right),
\end{aligned} \tag{3.134}$$

where (o) follows from (3.114).

Chapter 4

Sample Complexity Upper Bounds for Identification of Kronecker-structured Dictionaries

This chapter 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 chapter 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 errors $\{\varepsilon_k\}$ is shown to be $\max_{k\in[K]} \mathcal{O}(m_k p_k^3 \varepsilon_k^{-2})$.[1]

4.1 Introduction

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. 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. In this chapter, 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

[1]The results presented in this chapter have been published in Proceedings of 2017 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing [86] and IEEE Journal of Selected Topics in Signal Processing [12].

KS dictionary. To the best of our knowledge, this is the first work presenting such identification results for the KS-DL problem.

4.1.1 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, $\mathbf{D}_k^0 \in \mathbb{R}^{m_k \times p_k}, k \in [K]$, 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 $\varepsilon_k, k \in [K]$.

4.1.2 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 [14–16, 62, 69, 70].

In the case of unstructured dictionaries, several works do provide analytical results for the dictionary identifiability problem [1, 24, 42, 46, 47, 50, 58, 76]. 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 chapter, we focus on the Frobenius norm as the distance metric. Gribonval et al. [24] and Jung et al. [1] also consider this metric, with the latter work providing minimax lower bounds for dictionary reconstruction error. In particular, Jung et al. [1] 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. [24] 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 Chapter 3, we have obtained lower bounds on the minimax risk of KS DL for 2nd-order [65] and Kth-order tensors [66, 67], 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. [24], 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 assumption 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 [28, 67].

The rest of the chapter is organized as follows. We formulate the KS-DL problem in Section 4.2. In Section 4.3, we provide analysis for asymptotic recovery of coordinate dictionaries composing the KS dictionary and in Section 4.4, we present sample complexity results for identification of coordinate dictionaries that are based on the results of Section 4.3. Finally, we conclude the chapter in Section 4.5. In order to keep the main exposition simple, proofs of the lemmas and propositions are relegated to the appendix.

4.2 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 $\mathbf{y} \triangleq \text{vec}(\underline{\mathbf{Y}})$ using (1.10) as[2]

$$\mathbf{y} = \Big(\bigotimes_{k \in [K]} \mathbf{D}_k^0 \Big) \mathbf{x} + \mathbf{w}, \quad \|\mathbf{x}\|_0 \leq s, \tag{4.1}$$

where $\mathbf{x} = \text{vec}(\underline{\mathbf{X}}) \in \mathbb{R}^p$ 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 $\mathbf{w} = \text{vec}(\underline{\mathbf{W}}) \in \mathbb{R}^m$ 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, $\mathbf{x} = \text{vec}(\underline{\mathbf{X}})$ 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 [28].[4]

---

[2] We have reindexed $\mathbf{D}_k$'s in (1.10) 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 [24].

[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 $\{\mathcal{J} \subseteq [p] : |\mathcal{J}| = s\}$.

We now make the same assumptions for the distribution of $\mathbf{x}$ as assumptions A and B in Gribonval et al. [24]. 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 [24]).

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

Our goal in this chapter 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

$$f_{\mathbf{y}}\left(\mathbf{D}_{1:K}\right) \triangleq \inf_{\mathbf{x}'\in\mathbb{R}^p}\left\{\frac{1}{2}\left\|\mathbf{y} - \left(\bigotimes \mathbf{D}_k\right)\mathbf{x}'\right\|_2^2 + \lambda\|\mathbf{x}'\|_1\right\}, \text{and}$$
$$F_{\mathbf{Y}}\left(\mathbf{D}_{1:K}\right) \triangleq \frac{1}{N}\sum_{n=1}^{N} f_{\mathbf{y}_n}\left(\mathbf{D}_{1:K}\right), \tag{4.2}$$

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

$$\min_{\substack{\mathbf{D}_k\in\mathcal{D}_k \\ k\in[K]}} F_{\mathbf{Y}}\left(\mathbf{D}_{1:K}\right). \tag{4.3}$$

More specifically, given desired errors $\{\varepsilon_k\}_{k=1}^K$, we want a local minimum of (4.3) 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

[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 overcome this issue by defining distance metrics that capture the distance between these equivalent classes [46, 47, 76].

problem, we first minimize the statistical risk:

$$\min_{\substack{\mathbf{D}_k \in \mathcal{D}_k \\ k \in [K]}} f_{\mathbb{P}}\left(\mathbf{D}_{1:K}\right) \triangleq \min_{\substack{\mathbf{D}_k \in \mathcal{D}_k \\ k \in [K]}} \mathbb{E}_{\mathbf{y}}\left\{f_{\mathbf{y}}\left(\mathbf{D}_{1:K}\right)\right\}. \tag{4.4}$$

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 [24]).

4.3 Asympototic Identifiability Results

In this section, we provide an identifiability result for the KS-DL objective function in (4.4). The implications of this theorem are discussed in Section 4.5.

Theorem 4.1. *Suppose the observations are generated according to* (4.1) *and the dictionary coefficients follow the separable sparsity model of Section 4.2. Further, assume the following conditions are satisfied:*

$$s_k \leq \frac{p_k}{8\left(\left\|\mathbf{D}_k^0\right\|_2 + 1\right)^2}, \quad \max_{k \in [K]}\left\{\mu_{s_k}(\mathbf{D}_k^0)\right\} \leq \frac{1}{4}, \quad \mu_s(\mathbf{D}^0) < \frac{1}{2}, \tag{4.5}$$

and

$$\frac{\mathbb{E}\left\{x^2\right\}}{M_x \mathbb{E}\left\{|x|\right\}} > \frac{24\sqrt{3}(4.5^{K/2})K}{(1 - 2\mu_s(\mathbf{D}^0))} \max_{k \in [K]}\left\{\frac{s_k}{p_k}\left\|\mathbf{D}_k^{0\top}\mathbf{D}_k^0 - \mathbf{I}\right\|_F\left(\left\|\mathbf{D}_k^0\right\|_2 + 1\right)\right\}. \tag{4.6}$$

Define

$$\begin{aligned} C_{k,\min} &\triangleq 8(3^{\frac{K+1}{2}})\kappa_x^2\left(\frac{s_k}{p_k}\right)\left\|\mathbf{D}_k^{0\top}\mathbf{D}_k^0 - \mathbf{I}\right\|_F\left(\left\|\mathbf{D}_k^0\right\|_2 + 1\right), \\ C_{\max} &\triangleq \frac{1}{3K(1.5)^{K/2}}\frac{\mathbb{E}\left\{|x|\right\}}{M_x}(1 - 2\mu_s(\mathbf{D}^0)). \end{aligned} \tag{4.7}$$

Then, the map $\mathbf{D}_{1:K} \mapsto f_{\mathbb{P}}(\mathbf{D}_{1:K})$ *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*

$$\lambda \leq \frac{x_{\min}}{8 \times 3^{(K-1)/2}}, \tag{4.8}$$

$$\frac{\lambda C_{k,\min}}{\mathbb{E}\{|x|\}} < \varepsilon_k < \frac{\lambda C_{\max}}{\mathbb{E}\{|x|\}},\ k \in [K], \tag{4.9}$$

and

$$\frac{M_w}{M_x} < 3(1.5)^{K/2}\left(\frac{\lambda K C_{\max}}{\mathbb{E}\{|x|\}} - \sum_{k\in[K]} \varepsilon_k\right). \tag{4.10}$$

4.3.1 Discussion

Theorem 4.1 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 $f_{\mathbb{P}}(\mathbf{D}_{1:K})$ 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 [24, Theorem 1].

We now explicitly compare conditions in Theorem 4.1 with the corresponding ones for vectorized observations [24, Theorem 1]. Given that the coefficients are drawn from the separable sparsity model, the sparsity constraints for the coordinate dictionaries in (4.5) translate into

$$\frac{s}{p} = \prod_{k\in[K]} \frac{s_k}{p_k} \leq \frac{1}{8^K \prod_k \left(\left\|\mathbf{D}_k^0\right\|_2 + 1\right)^2}. \tag{4.11}$$

Therefore, we have $\frac{s}{p} = \mathcal{O}\left(\frac{1}{\prod_k \left\|\mathbf{D}_k^0\right\|_2^2}\right) = \mathcal{O}\left(\frac{1}{\|\mathbf{D}^0\|_2^2}\right)$. Using the fact that $\left\|\mathbf{D}^0\right\|_2 \geq \|\mathbf{D}^0\|_F/\sqrt{m} = \sqrt{p}/\sqrt{m}$, this translates into sparsity order $s = \mathcal{O}(m)$. Next, the left hand side of the condition in (4.6) is less than 1. Moreover, from properties of the Frobenius norm, it is easy to show that $\left\|\mathbf{D}_k^{0\top}\mathbf{D}_k^0 - \mathbf{I}\right\|_F \geq \sqrt{p_k(p_k - m_k)/m_k}$. 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 (4.6) is lower bounded by $\Omega\left(\max_k s_k\sqrt{(p_k - m_k)/m_k^2}\right)$. Therefore, Theorem 4.1 applies to coordinate dictionaries with dimensions $p_k \leq m_k^2$ and subsequently, KS dictionaries with $p \leq m^2$. Both the sparsity order and dictionary dimensions are in line with the scaling results for vectorized data [24].

4.3.2 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 $\mathbf{D}_k, \mathbf{D}_k' \in \mathbb{R}^{m_k \times p_k}, k \in [K]$ [87]. 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

$$\Delta f_{\mathbb{P}}(r_{1:K}) \triangleq \inf_{\mathbf{D}_k \in \mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)} \Delta f_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0\right) > 0. \tag{4.12}$$

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 $\Delta f_{\mathbb{P}}(r_{1:K}) > 0$, 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

$$\mathbf{D}_{\mathcal{J}} = \bigotimes \mathbf{D}_{k,\mathcal{J}_k}, \tag{4.13}$$

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

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

$f_{\mathbf{y}}\left(\mathbf{D}_{1:K}\right)$ given $\widehat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}$, which we denote as $\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right)$. Defining

$$\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right) \triangleq \mathbb{E}\left\{\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right)\right\}, \tag{4.14}$$

we expand $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)$ using (4.13) and separate the terms that depend on each radius $\varepsilon_k = \|\mathbf{D}_k - \mathbf{D}^0_k\|_F$ to obtain conditions for sparsity levels $s_k, k \in [K]$, and coordinate dictionaries such that $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) > 0$. Finally, we derive conditions on M_w, coordinate dictionary coherences and ε_k's that ensure $\widehat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}$ and $\Delta f_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}\right) = \Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)$.

Remark 4.1. The key assumption in the proof of Theorem 4.1 is expanding $\mathbf{D}_{\mathcal{J}}$ according to (4.13). This is a consequence of the separable sparsity model for dictionary coefficients.

Remark 4.2. Although some of the forthcoming lemmas needed of Theorem 4.1 impose conditions on $\mathbf{D}_k$'s as well as true coordinate dictionaries $\mathbf{D}^0_k$'s, we later translate these conditions exclusively in terms of $\mathbf{D}^0_k$'s and ε_k's.

The proof of Theorem 4.1 relies on the following propositions and lemmas. The proofs of these are provided in Appendix A.

Proposition 4.1. *Suppose the following inequalities hold for $k \in [K]$:*

$$s_k \leq \frac{p_k}{8(\|\mathbf{D}^0_k\|_2 + 1)^2} \quad \textit{and} \quad \max_{k\in[K]} \left\{\delta_{s_k}(\mathbf{D}^0_k)\right\} \leq \frac{1}{4}. \tag{4.15}$$

Then, for

$$\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\left\{|x|\right\}} \leq \frac{1}{8 \times 3^{(K-1)/2}}, \tag{4.16}$$

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

$$\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) \geq \frac{s\mathbb{E}\{x^2\}}{8} \sum_{k\in[K]} \frac{\varepsilon_k}{p_k}\left(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda})\right), \tag{4.17}$$

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

$$\bar{\lambda} \leq \frac{0.15}{\max_{k\in[K]} C_{k,\min}}, \tag{4.18}$$

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

The proof of Proposition 4.1 relies on the following lemmas as well as supporting lemmas from the analysis of vectorized data [24, Lemmas 4,6,7,15,16].

Lemma 4.1. *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]

$$\mathbf{P}_{\mathbf{D}_{\mathcal{J}}} = \bigotimes \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}, \mathbf{D}^+_{\mathcal{J}} = \bigotimes \mathbf{D}^+_{k,\mathcal{J}_k}, \mathbf{H}_{\mathbf{D}_{\mathcal{J}}} = \bigotimes \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}, \tag{4.19}$$

where $\mathbf{P}$ *and* $\mathbf{H}$ *are defined in Section 1.3.*

Lemma 4.2. *Given* $\mathbf{D}_{1:K}$ *and* $\mathbf{D}^0_{1:K}$, *the difference*

$$\bigotimes \mathbf{D}_k - \bigotimes \mathbf{D}^0_k = \sum_{k\in[K]} \widetilde{\mathbf{D}}_{k,1} \otimes \cdots \otimes \left(\mathbf{D}_k - \mathbf{D}^0_k\right) \otimes \cdots \otimes \widetilde{\mathbf{D}}_{k,K}, \tag{4.20}$$

where without loss of generality, each $\widetilde{\mathbf{D}}_{k,i}$ *is equal to either* $\mathbf{D}^0_i$ *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 chapter.

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

Lemma 4.3. *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]

$$\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right) \triangleq \inf_{\substack{\mathbf{x}\in\mathbb{R}^p \\ \mathrm{supp}(\mathbf{x})\subset\mathcal{J}}} \frac{1}{2}\left\|\mathbf{y} - \left(\bigotimes \mathbf{D}_k\right)\mathbf{x}\right\|_2^2 + \lambda\boldsymbol{\sigma}^\top\mathbf{x}. \tag{4.21}$$

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

$$\widehat{\mathbf{x}}_{\mathcal{J}} = \left(\bigotimes \mathbf{D}_{k,\mathcal{J}_k}^+\right)\mathbf{y} - \lambda\left(\bigotimes \left(\mathbf{D}_{k,\mathcal{J}_k}^\top \mathbf{D}_{k,\mathcal{J}_k}\right)^{-1}\right)\boldsymbol{\sigma}_{\mathcal{J}}, \tag{4.22}$$

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:

$$\begin{aligned}\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right) &= \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\mathbf{y}^\top\left(\bigotimes \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\mathbf{y} \\ &+ \lambda\boldsymbol{\sigma}_{\mathcal{J}}^\top\left(\bigotimes \mathbf{D}_{k,\mathcal{J}_k}^+\right)\mathbf{y} - \frac{\lambda^2}{2}\boldsymbol{\sigma}_{\mathcal{J}}^\top\left(\bigotimes \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\boldsymbol{\sigma}_{\mathcal{J}}.\end{aligned} \tag{4.23}$$

Lemma 4.4. *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*

$$\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0\middle|\boldsymbol{\sigma}\right) \triangleq \phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right) - \phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right), \tag{4.24}$$

we have

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

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

$$\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{H}_{\mathbf{D}^0_{k,\mathcal{J}_k}}-\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right]\right\}\ldots\mathbb{E}_{\mathcal{J}_K}\left\{\mathrm{Tr}\left[\mathbf{H}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\right]\right\}. \tag{4.25}$$

Lemma 4.5. *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:*

$$\|\mathbf{D}_{k,\mathcal{J}_k}\|_2 = \left\|{\mathbf{D}_{k,\mathcal{J}_k}}^\top\right\|_2 \leq \sqrt{1+\delta_{s_k}(\mathbf{D}_k)}, \tag{4.26}$$

$$\delta_{s_k}(\mathbf{D}_k) \leq \mu_{s_k-1}(\mathbf{D}_k). \tag{4.27}$$

Lemma 4.6 (Lemma 4 [24])**.** *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*

$$\left\|\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right\|_2 \leq \frac{1}{1-\delta_{s_k}(\mathbf{D}_k)}, \quad \left\|\mathbf{D}^+_{k,\mathcal{J}_k}\right\|_2 \leq \frac{1}{\sqrt{1-\delta_{s_k}(\mathbf{D}_k)}}, \tag{4.28}$$

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

$$1-\delta_{s_k}(\mathbf{D}'_k) \geq (\sqrt{1-\delta_{s_k}(\mathbf{D}_k)}-\varepsilon_k)^2 \triangleq 1-\delta_k. \tag{4.29}$$

Lemma 4.7 (Lemma 6 [24])**.** *Given any $\mathbf{D}^1_k, \mathbf{D}^2_k \in \mathcal{D}_k$, there exist $\mathbf{V}_k \in \mathbb{R}^{m_k \times p_k}$ with $\mathrm{diag}\left({\mathbf{D}^1_k}^\top \mathbf{V}_k\right) = \mathbf{0}$ and $\mathrm{diag}\left(\mathbf{V}_k^\top \mathbf{V}_k\right) = \mathbf{I}_{p_k}$ and a vector $\boldsymbol{\theta}_k \triangleq \boldsymbol{\theta}_k(\mathbf{D}^1_k, \mathbf{D}^2_k) \in [0,\pi]^{p_k}$, such that*

$$\mathbf{D}^2_k = \mathbf{D}^1_k \mathbf{C}_k(\boldsymbol{\theta}_k) + \mathbf{V}_k \mathbf{S}_k(\boldsymbol{\theta}_k), \tag{4.30}$$

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

$$\begin{aligned}
&\frac{2}{\pi}\theta_{k,j} \leq \|\mathbf{d}^2_{k,j} - \mathbf{d}^1_{k,j}\|_2 = 2\sin\left(\frac{\theta_{k,j}}{2}\right) \leq \theta_{k,j}, \text{ and} \\
&\frac{2}{\pi}\|\boldsymbol{\theta}_k\|_2 \leq \|\mathbf{D}^2_k - \mathbf{D}^1_k\|_F \leq \|\boldsymbol{\theta}_k\|_2,
\end{aligned} \tag{4.31}$$

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

Lemma 4.8. *Fix* $\mathbf{D}_{1:K}$ *and* $\mathbf{D}^0_{1:K}$*, and suppose* $\{A_k\}, \{B_k\}, \{\delta_k\}$ *satisfy the following:*

$$\begin{aligned} &A_k \geq \max\left\{\|\mathbf{D}_k^\top \mathbf{D}_k - \mathbf{I}_{p_k}\|_F, \|{\mathbf{D}_k^0}^\top \mathbf{D}_k^0 - \mathbf{I}_{p_k}\|_F\right\}, \\ &B_k \geq \max\left\{\|\mathbf{D}_k\|_2, \|\mathbf{D}_k^0\|_2\right\}, \textit{and} \\ &\delta_k \geq \max\left\{\delta_{s_k}(\mathbf{D}_k), \delta_{s_k}(\mathbf{D}_k^0)\right\}. \end{aligned} \tag{4.32}$$

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

$$\begin{aligned} \Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) \geq \frac{s\mathbb{E}\{x^2\}}{2} \sum_{k\in[K]} \frac{\|\boldsymbol{\theta}_k\|_2}{p_k}\Bigg[\|\boldsymbol{\theta}_k\|_2\left(1 - \frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k} - \bar{\lambda}\kappa_x^2\delta_{-k}\right) \\ - \left(\delta_{-k} + 2\bar{\lambda}\prod_{i\in[K]}\frac{1}{1-\delta_i}\right)\bar{\lambda}\kappa_x^2\frac{s_k}{p_k}\frac{2A_kB_k}{1-\delta_k}\Bigg], \end{aligned} \tag{4.33}$$

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

Proposition 4.1 shows $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) > 0$. However, given $\widehat{\mathbf{x}}$, the solution of $\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right)$, $\widehat{\boldsymbol{\sigma}} = \operatorname{sign}\left(\widehat{\mathbf{x}}\right)$ 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 4.2. *Let the generating coordinate dictionaries* $\{\mathbf{D}_k^0 \in \mathcal{D}_k\}$ *satisfy:*

$$\mu_s(\mathbf{D}^0) < \frac{1}{2}, \quad \max_k\{\delta_{s_k}(\mathbf{D}_k^0)\} < \frac{1}{4}. \tag{4.34}$$

Suppose $\bar{\lambda} = \dfrac{\lambda}{\mathbb{E}\{|x|\}} \leq \dfrac{x_{\min}}{2\mathbb{E}\{|x|\}}$ *and*

$$\max_{k\in[K]}\{\varepsilon_k\} \leq \min\left\{\bar{\lambda}C_{\max}, 0.15\right\}. \tag{4.35}$$

If the following is satisfied:

$$\frac{M_w}{M_x} < 3(1.5)^{K/2}\left(\bar{\lambda}KC_{\max} - \sum_{k\in[K]}\varepsilon_k\right), \tag{4.36}$$

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* (4.22)

is almost surely the minimizer of the map $\mathbf{x}' \mapsto \frac{1}{2}\left\|\mathbf{y} - \left(\bigotimes \mathbf{D}_k\right)\mathbf{x}'\right\|_2^2 + \lambda\|\mathbf{x}'\|_1$ *and* $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) = \Delta f_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0\right)$.

Remark 4.3. Note that $\mu_s(\mathbf{D}^0) < \frac{1}{2}$ in (4.34) can be satisfied by ensuring that the right hand side of (1.4) 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 4.2 relies on the following lemmas and [24, Lemmas 10–13].

Lemma 4.9 (Lemma 13 [24])**.** *Assume* $\mu_s(\mathbf{D}) < \dfrac{1}{2}$. *If*

$$\min_{j\in\mathcal{J}} |x_j| \geq 2\lambda, \ \textit{and} \ \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 < \lambda(1 - 2\mu_s(\mathbf{D})) \tag{4.37}$$

hold for generating $\mathbf{x}$, *then* $\widehat{\mathbf{x}}$ *defined in* (4.22) *is the unique solution of* $\min_{\mathbf{x}'} \frac{1}{2}\left\|\mathbf{y} - \left(\bigotimes \mathbf{D}_k\right)\mathbf{x}'\right\|_2 + \lambda\|\mathbf{x}'\|_1$.

Lemma 4.10. *For any* $\mathbf{D}^0 = \bigotimes \mathbf{D}_k^0$ *and* $\mathbf{D} = \bigotimes \mathbf{D}_k$ *such that* $\mathbf{D}_k \in \bar{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0)$, *for* $k \in [K]$, *suppose the following inequalities are satisfied:*

$$\max_{k\in[K]}\{\delta_{s_k}(\mathbf{D}_k^0)\} \leq \frac{1}{4}, \quad \textit{and} \quad \max_{k\in[K]} \varepsilon_k \leq 0.15. \tag{4.38}$$

Then, we have

$$\mu_s(\mathbf{D}) \leq \mu_s(\mathbf{D}^0) + 2(1.5)^{K/2}\sqrt{s}\left(\sum_{k\in[K]} \varepsilon_k\right). \tag{4.39}$$

Proof of Theorem 4.1. To prove this theorem, we use Proposition 4.1 to show that $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) > 0$, and then use Proposition 4.2 to show that $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) = \Delta f_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0\right)$. The assumptions in (4.5) ensure that the conditions in (4.15) and (4.34) are satisfied for Proposition 4.1 and Proposition 4.2, respectively. Assumptions (4.6) and (4.8) ensure that the conditions in (4.16) and (4.18) are satisfied for Proposition 4.1, $\bar{\lambda} \leq \dfrac{x_{\min}}{2\mathbb{E}\{|x|\}}$ holds for Proposition 4.2, and $\max_{k\in[K]}\{C_{k,\min}\} < C_{\max}$. Hence, according to Proposition 4.1, $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) > 0$ for all $\varepsilon_k \in (\bar{\lambda}C_{k,\min}, 0.15], k \in [K]$. Finally, using the assumption in (4.10) implies $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) = \Delta f_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0\right)$ for all $\varepsilon_k \leq \bar{\lambda}C_{\max}, k \in [K]$. Furthermore, the assumption in (4.8) implies $C_{\max}\bar{\lambda} \leq 0.15$.

Consequently, for any $\{\varepsilon_k > 0, k \in [K]\}$ satisfying the conditions in (4.9), $\mathbf{D}_{1:K} \to f_{\mathbb{P}}(\mathbf{D}_{1:K})$ admits a local minimum $\widehat{\mathbf{D}} = \bigotimes \widehat{\mathbf{D}}_k$ such that $\widehat{\mathbf{D}}_k \in \mathcal{B}_{\varepsilon_k}(\mathbf{D}_k^0), k \in [K]$. $\square$

4.4 Finite Sample Identifiability Results

We now focus on leveraging Theorem 4.1 and solving (4.3) to derive finite-sample bounds for KS dictionary identifiability. Compared to Gribonval et al. [24], 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 4.2. *Suppose the observations are generated according to* (4.1) *and the dictionary coefficients follow the separable sparsity model of Section 4.2 such that* (4.5) *to* (4.10) *are satisfied. Next, fix any* $\xi \in (0, \infty)$*. Then, for any number of observations satisfying*

$$N = \max_{k \in [K]} \Omega\left(\frac{p_k^2(\xi + m_k p_k)}{(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda}))^2} \left(\frac{2^K (1 + \bar{\lambda}^2) M_x^2}{s^2 \mathbb{E}\{x^2\}^2} + \left(\frac{M_w}{s \mathbb{E}\{x^2\}} \right)^2 \right) \right), \tag{4.40}$$

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* $k \in [K]$*.*

4.4.1 Discussion

Let us make some remarks about implications of Theorem 4.2. First, sample complexity has an inverse relationship with signal to noise ratio (SNR),[9] defined as

$$\text{SNR} \triangleq \frac{\mathbb{E}\{\|\mathbf{x}\|_2^2\}}{\mathbb{E}\{\|\mathbf{w}\|_2^2\}} = \frac{s \mathbb{E}\{x^2\}}{m \mathbb{E}\{w^2\}}. \tag{4.41}$$

Looking at the terms on the right hand side of (4.40) in Theorem 4.2, $M_x/(s\mathbb{E}\left\{x^2\right\})$ is related to the deviation of $\|\mathbf{x}\|_2$ from its mean, $\mathbb{E}\left\{\|\mathbf{x}\|_2\right\}$, and depends on the coefficient distribution, while $M_w/(s\mathbb{E}\left\{x^2\right\})$ is related to $1/\,\text{SNR}$ and depends on the noise and coefficient distributions.

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

Table 4.1: 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	$\frac{mp^2}{\varepsilon^2}$ [1]	$\frac{p\sum_k m_k p_k}{\varepsilon^2}$
Achievability Bound	$\frac{mp^3}{\varepsilon^2}$ [24]	$\max_k \frac{m_k p_k^3}{\varepsilon_k^2}$

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 (4.40) is proportional to $\max_k \varepsilon_k^{-2}$. We note that the sample complexity results obtained in [24] that are independent of $\varepsilon \triangleq \left\|\mathbf{D}-\mathbf{D}^0\right\|_F$ only hold for the noiseless setting and the dependency on ε^{-2} is inevitable for noisy observations [24]. Furthermore, given the condition on the range of ε_k's in (4.9), 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 (4.40) 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 [24], 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 Corollary 3.4, 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

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

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

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

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 4.1 for a fixed K.

4.4.2 Proof Outline

We follow a similar approach used in [24, Theorem 2] for vectorized data. We show that, with high probability,

$$\Delta F_{\mathbf{Y}}(r_{1:K}) \triangleq \inf_{\mathbf{D}_k \in \mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)} \Delta F_{\mathbf{Y}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0\right) \tag{4.42}$$

converges uniformly to its expectation,

$$\Delta f_{\mathbb{P}}(r_{1:K}) \triangleq \inf_{\mathbf{D}_k \in \mathcal{S}_{\varepsilon_k}(\mathbf{D}_k^0)} \Delta f_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0\right). \tag{4.43}$$

In other words, with high probability,

$$|\Delta F_{\mathbf{Y}}(r_{1:K}) - \Delta f_{\mathbb{P}}(r_{1:K})| \leq \eta_N, \tag{4.44}$$

where η_N is a parameter that depends on the probability and other parameters in the problem. This implies $\Delta F_{\mathbf{Y}}(r_{1:K}) \geq \Delta f_{\mathbb{P}}(r_{1:K}) - 2\eta_N$. In Theorem 4.1, we obtained conditions that ensure $\Delta f_{\mathbb{P}}(r_{1:K}) > 0$. Thus, if $2\eta_N < \Delta f_{\mathbb{P}}(r_{1:K})$ is satisfied, this implies $\Delta F_{\mathbf{Y}}(r_{1:K}) > 0$, and we can use arguments similar to the proof of Theorem 4.1 to show that $\mathbf{D}_{1:K} \mapsto F_{\mathbf{Y}}\left(\mathbf{D}_{1:K}\right)$ 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 4.1, 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 (4.17) and obtain conditions on the sufficient number of samples based on each coordinate dictionary dimension and recovery error.

The proof of Theorem 4.2 relies on the following definition and lemmas. The proofs of these are provided in Appendix B.

Definition 4.3 (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 $\{\mathbf{D}_k, \mathbf{D}'_k \in \mathcal{D}_k\}_{k=1}^K$:

$$\left|f\left(\mathbf{D}_{1:K}\right) - f\left(\mathbf{D}'_{1:K}\right)\right| \leq \sum_{k\in[K]} L_k \left\|\mathbf{D}_k - \mathbf{D}'_k\right\|_F . \tag{4.45}$$

Lemma 4.11 (Rademacher averages [24])**.** *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 $1 - e^{-\xi}$:*

$$\begin{aligned}&\sup_{f\in\mathcal{F}}\left(\frac{1}{N}\sum_{n\in[N]} f\left(X_n\right) - \mathbb{E}_X\left\{f\left(X\right)\right\}\right)\\ &\quad\leq 2\sqrt{\frac{\pi}{2}}\mathbb{E}_{X,\beta_{1:N}}\left\{\sup_{f\in\mathcal{F}}\left(\frac{1}{N}\sum_{n\in[N]}\beta_n f\left(X_n\right)\right)\right\} + B\sqrt{\frac{2\xi}{N}},\end{aligned} \tag{4.46}$$

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

Lemma 4.12. *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:*

$$\begin{aligned}&\sup_{\substack{\mathbf{D}_k\in\overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0)\\ k\in[K]}}\left|\frac{1}{N}\sum_{n\in[N]} h_n(\mathbf{D}_{1:K}) - \mathbb{E}\left\{h(\mathbf{D}_{1:K})\right\}\right|\\ &\qquad\leq 4\sqrt{\frac{\pi}{2N}}\left(\sum_{k\in[K]} L_k\varepsilon_k\sqrt{Km_kp_k}\right) + B\sqrt{\frac{2\xi}{N}}.\end{aligned} \tag{4.47}$$

Lemma 4.13 (Lemma 5 [24]). *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*

$$\begin{aligned}
&\|\mathbf{I} - \mathbf{D}^+_{k,\mathcal{J}_k}\mathbf{D}'_{k,\mathcal{J}_k}\|_2 \leq (1-\delta_k)^{-1/2}\|\mathbf{D}_k - \mathbf{D}'_k\|_F,\\
&\|\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}} - \mathbf{H}_{\mathbf{D}'_{k,\mathcal{J}_k}}\|_2 \leq 2(1-\delta_k)^{-3/2}\|\mathbf{D}_k - \mathbf{D}'_k\|_F,\\
&\|\mathbf{D}^+_{k,\mathcal{J}_k} - \mathbf{D}'^+_{k,\mathcal{J}_k}\|_2 \leq 2(1-\delta_k)^{-1}\|\mathbf{D}_k - \mathbf{D}'_k\|_F, \text{ and}\\
&\|\mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}'_{k,\mathcal{J}_k}}\|_2 \leq 2(1-\delta_k)^{-1/2}\|\mathbf{D}_k - \mathbf{D}'_k\|_F. \qquad (4.48)
\end{aligned}$$

Lemma 4.14. *Consider $\mathbf{D}^0_k \in \mathcal{D}_k$ and ε_k's such that $\varepsilon_k < \sqrt{1-\delta_{s_k}(\mathbf{D}^0_k)}$, for $k \in [K]$ and define $\sqrt{1-\delta_k} \triangleq \sqrt{1-\delta_{s_k}(\mathbf{D}^0_k)} - \varepsilon_k > 0$. The function $\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)$ is almost surely coordinate-wise Lipschitz continuous on $\left\{\mathcal{B}_{\varepsilon_k}(\mathbf{D}^0_k)\right\}_{k=1}^K$ with Lipschitz constants*

$$L_k \triangleq (1-\delta_k)^{-1/2}\left(M_x\left(\prod_{k\in[K]}\sqrt{1+\delta_{s_k}(\mathbf{D}^0_k)}\right) + M_w + \lambda\sqrt{s}\prod_{k\in[K]}(1-\delta_k)^{-1/2}\right)^2, \qquad (4.49)$$

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

Proof of Theorem 2. From Lemmas 4.12 and 4.14, we have that with probability at least $1 - e^{-\xi}$:

$$\begin{aligned}
&\sup_{\substack{\mathbf{D}_k \in \overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}^0_k)\\ k\in[K]}} \left|\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) - \Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right|\\
&\qquad \leq \sqrt{\frac{2}{N}}\sum_{k\in[K]} L_k\varepsilon_k\left(2\sqrt{\pi m_k p_k} + \sqrt{\xi}\right), \qquad (4.50)
\end{aligned}$$

where L_k is defined in (4.49). From (4.50), we obtain

$$\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) > \Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) - 2\eta_N, \qquad (4.51)$$

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 4.1, we derived conditions that ensure $\Delta f_{\mathbf{y}}(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}) = \Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)$ and $\Delta f_{\mathbb{P}}(\mathbf{D}_{1:K}; \mathbf{D}^0_{1:K}) =$

$\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)$. Therefore, given that the conditions in Theorem 4.1 are satisfied, $\Delta F_{\mathbf{Y}}(r_{1:K}) > \Delta f_{\mathbb{P}}(r_{1:K}) - 2\eta_N$, and the existence of a local minimum of $F_{\mathbf{Y}}(\mathbf{D}_{1:K})$ within radii ε_k around $\mathbf{D}^0_k$, $k \in [K]$, is guaranteed with probability at least $1 - e^{-\xi}$ as soon as $2\eta_N < \Delta f_{\mathbb{P}}(r_{1:K})$. According to (4.17),
$\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) \geq \frac{s\mathbb{E}\{x^2\}}{8}\sum_{k\in[K]}\frac{\varepsilon_k}{p_k}\left(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda})\right)$; 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

$$N_k = \left(2\sqrt{\pi m_k p_k} + \sqrt{\xi}\right)^2\left(\frac{2^{4.5}L_k p_k}{s\mathbb{E}\{x^2\}(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda}))}\right)^2. \tag{4.52}$$

Furthermore, we can upper bound L_k by

$$\begin{aligned} L_k &\overset{(a)}{\leq} \sqrt{2}\left(1.25^{K/2}M_x + M_w + 2^{K/2}\lambda\sqrt{s}\right)^2 \\ &\overset{(b)}{\leq} \sqrt{2}c_1\left(\left(1.25^K + 2^K\bar{\lambda}^2\right)M_x^2 + M_w^2\right), \end{aligned} \tag{4.53}$$

where c_1 is some positive constant, (a) follows from the fact that given the assumption in (4.15), assumptions in Lemma 4.14 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:

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

Substituting (4.53) in (4.52) 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 4.4. 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 [87, Theorem 1] that prove uniform convergence results using covering number arguments for various classes of dictionaries. In this case, we get $\eta_N \leq c\sqrt{\frac{(\sum_k m_k p_k + \xi)\log N}{N}}$ for some constant c, where an extra $\sqrt{\log N}$ term appears compared to (4.47). Therefore, Lemma 4.12 provides a tighter upper bound.

4.5 Conclusion

In this chapter, 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 4.2.

4.6 Appendix

4.6.1 Proof of Lemma 4.2

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

$$\begin{aligned}(\mathbf{D}_1 \otimes \mathbf{D}_2) - \left(\mathbf{D}_1^0 \otimes \mathbf{D}_2^0\right) &= \left(\mathbf{D}_1 - \mathbf{D}_1^0\right) \otimes \mathbf{D}_2^0 + \mathbf{D}_1 \otimes \left(\mathbf{D}_2 - \mathbf{D}_2^0\right) \\ &= \left(\mathbf{D}_1 - \mathbf{D}_1^0\right) \otimes \mathbf{D}_2 + \mathbf{D}_1^0 \otimes \left(\mathbf{D}_2 - \mathbf{D}_2^0\right). \end{aligned} \tag{4.54}$$

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

$$\bigotimes_{k\in[K]} \mathbf{D}_k - \bigotimes_{k\in[K]} \mathbf{D}_k^0 = \sum_{k\in[K]} \widetilde{\mathbf{D}}_{k,1} \otimes \cdots \otimes \left(\mathbf{D}_k - \mathbf{D}_k^0\right) \otimes \cdots \otimes \widetilde{\mathbf{D}}_{k,K}. \tag{4.55}$$

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

$$\begin{aligned}
\bigotimes_{k\in[K+1]} \mathbf{D}_k - \bigotimes_{k\in[K+1]} \mathbf{D}_k^0 &= \left(\bigotimes_{k\in[K]} \mathbf{D}_k\right) \otimes \mathbf{D}_{K+1} - \left(\bigotimes_{k\in[K]} \mathbf{D}_k^0\right) \otimes \mathbf{D}_{K+1}^0 \\
&\overset{(a)}{=} \left(\bigotimes_{k\in[K]} \mathbf{D}_k - \bigotimes_{k\in[K]} \mathbf{D}_k^0\right) \otimes \mathbf{D}_{K+1}^0 \\
&\quad + \left(\bigotimes_{k\in[K]} \mathbf{D}_k\right) \left(\mathbf{D}_{K+1} - \mathbf{D}_{K+1}^0\right) \\
&\overset{(b)}{=} \left(\sum_{k\in[K]} \widetilde{\mathbf{D}}_{k,1} \otimes \cdots \otimes \left(\mathbf{D}_k - \mathbf{D}_k^0\right) \otimes \cdots \otimes \widetilde{\mathbf{D}}_{k,K}\right) \\
&\qquad \otimes \mathbf{D}_{K+1}^0 + \left(\bigotimes_{k\in[K]} \mathbf{D}_k\right) \left(\mathbf{D}_{K+1} - \mathbf{D}_{K+1}^0\right) \\
&\overset{(c)}{=} \sum_{k\in[K+1]} \widetilde{\mathbf{D}}_{k,1} \otimes \cdots \otimes \left(\mathbf{D}_k - \mathbf{D}_k^0\right) \otimes \cdots \otimes \widetilde{\mathbf{D}}_{k,K+1},
\end{aligned} \tag{4.56}$$

where (a) follows from (4.54), (b) follows from (4.55) and (c) follows from replacing $\mathbf{D}_{K+1}^0$ 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.

4.6.2 Proof of Lemma 4.3

Using the same definition as Gribonval et al. [24, 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 (4.22) for $\widehat{\mathbf{x}}$. Substituting $\widehat{\mathbf{x}}$ in (4.21), we get

$$\begin{aligned}
\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right) &= \frac{1}{2}\Bigg[\|\mathbf{y}\|_2^2 - \left(\left(\bigotimes \mathbf{D}_{k,\mathcal{J}_k}^\top\right)\mathbf{y} - \lambda\boldsymbol{\sigma}_{\mathcal{J}}\right)^\top \\
&\quad \left(\bigotimes (\mathbf{D}_{k,\mathcal{J}_k}^\top \mathbf{D}_{k,\mathcal{J}_k})^{-1}\right)\left(\left(\bigotimes \mathbf{D}_{k,\mathcal{J}_k}^\top\right)\mathbf{y} - \lambda\boldsymbol{\sigma}_{\mathcal{J}}\right)\Bigg] \\
&\overset{(a)}{=} \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\mathbf{y}^\top\left(\bigotimes \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\mathbf{y} + \lambda\boldsymbol{\sigma}_{\mathcal{J}}^\top\left(\bigotimes \mathbf{D}_{k,\mathcal{J}_k}^+\right)\mathbf{y} - \frac{\lambda^2}{2}\boldsymbol{\sigma}_{\mathcal{J}}^\top\left(\bigotimes \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\boldsymbol{\sigma}_{\mathcal{J}},
\end{aligned} \tag{4.57}$$

where (a) follows from (4.19).

4.6.3 Proof of Lemma 4.4

We use the expression for $\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right)$ from (4.23). 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

$$\begin{aligned}
\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) &= \phi_{\mathbf{y}}\left(\mathbf{D}_{1:K}|\boldsymbol{\sigma}\right) - \phi_{\mathbf{y}}\left(\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) \\
&= \frac{1}{2}\mathbf{y}^\top\left(\bigotimes \mathbf{P}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \bigotimes \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\mathbf{y} - \lambda\boldsymbol{\sigma}_{\mathcal{J}}^\top\left(\bigotimes \mathbf{D}'^{+}_{k,\mathcal{J}_k} - \bigotimes \mathbf{D}^{+}_{k,\mathcal{J}_k}\right)\mathbf{y} \\
&\quad + \frac{\lambda^2}{2}\boldsymbol{\sigma}_{\mathcal{J}}^\top\left(\bigotimes \mathbf{H}_{\mathbf{D}'_{k,\mathcal{J}_k}} - \bigotimes \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\boldsymbol{\sigma}_{\mathcal{J}}. \qquad (4.58)
\end{aligned}$$

We substitute $\mathbf{y}=\left(\bigotimes \mathbf{D}^0_k\right)\mathbf{x}+\mathbf{w}=\left(\bigotimes \mathbf{D}^0_{k,\mathcal{J}_k}\right)\mathbf{x}_{\mathcal{J}}+\mathbf{w}$ and break up the sum in (4.58) into 6 terms:

$$\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) = \sum_{i\in[6]}\Delta\phi_i\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right), \qquad (4.59)$$

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\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 \\
&\qquad \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}_1^{0\top}\mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\mathbf{D}^0_1\right)\otimes\cdots\otimes \\
&\qquad \left(\mathbf{D}_k^{0\top}(\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{D}_K^{0\top}\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\mathbf{D}^0_K\right)\Bigg)\mathbf{x}, \\
\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 \\
&\qquad \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}, \\
\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 \\
&\qquad \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}$$

$$\Delta\phi_4\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) = -\lambda\boldsymbol{\sigma}_{\mathcal{J}}^{\top}\Bigg(\sum_{k\in[K]}\left(\widetilde{\mathbf{D}}^{+}_{1,\mathcal{J}_1}\mathbf{D}^{0}_{1}\right)\otimes\cdots\otimes$$
$$\left((\mathbf{D}'^{+}_{k,\mathcal{J}_k}-\mathbf{D}^{+}_{k,\mathcal{J}_k})\mathbf{D}^{0}_{k}\right)\otimes\cdots\otimes\left(\widetilde{\mathbf{D}}^{+}_{K,\mathcal{J}_K}\mathbf{D}^{0}_{K}\right)\Bigg)\mathbf{x},$$
$$\Delta\phi_5\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) = -\lambda\boldsymbol{\sigma}_{\mathcal{J}}^{\top}\Bigg(\sum_{k\in[K]}\widetilde{\mathbf{D}}^{+}_{1,\mathcal{J}_1}\otimes\cdots\otimes$$
$$\left(\mathbf{D}'^{+}_{k,\mathcal{J}_k}-\mathbf{D}^{+}_{k,\mathcal{J}_k}\right)\otimes\cdots\otimes\widetilde{\mathbf{D}}^{+}_{K,\mathcal{J}_K}\Bigg)\mathbf{w},\text{ and}$$
$$\Delta\phi_6\left(\mathbf{D}_{1:K};\mathbf{D}'_{1:K}|\boldsymbol{\sigma}\right) = \frac{\lambda^2}{2}\boldsymbol{\sigma}_{\mathcal{J}}^{\top}\Bigg(\sum_{k\in[K]}\mathbf{H}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}}\otimes\cdots\otimes$$
$$\left(\mathbf{H}_{\mathbf{D}'_{k,\mathcal{J}_k}}-\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\right)\otimes\cdots\otimes\mathbf{H}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\Bigg)\boldsymbol{\sigma}_{\mathcal{J}},\tag{4.60}$$

where (a) follows from Lemma 4.2 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:

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

$$= \frac{\lambda^2}{2} \operatorname{Tr}\left[\boldsymbol{\sigma}_{\mathcal{J}} \boldsymbol{\sigma}_{\mathcal{J}}^{\top} \Big(\sum_{k\in[K]} \mathbf{H}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}} \otimes \cdots \otimes \left(\mathbf{H}_{\mathbf{D}^0_{k,\mathcal{J}_k}} - \mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}} \right) \otimes \cdots \otimes \mathbf{H}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}} \Big) \right], \tag{4.61}$$

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 (4.61), we get

$$\begin{aligned}
&\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}} \Big\{ \sum_{k\in[K]} \operatorname{Tr}\left[\mathbf{D}^{0\top}_1 \mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}} \mathbf{D}^0_1\right] \dots \\
&\operatorname{Tr}\left[\mathbf{D}^{0\top}_k (\mathbf{I}_{m_k} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}) \mathbf{D}^0_k\right] \dots \operatorname{Tr}\left[\mathbf{D}^{0\top}_K \mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}} \mathbf{D}^0_K\right] \Big\} \\
&= \frac{\mathbb{E}\{x^2\}}{2} \sum_{k\in[K]} \mathbb{E}_{\mathcal{J}_1} \left\{ \operatorname{Tr}\left[\mathbf{D}^{0\top}_1 \mathbf{P}_{\widetilde{\mathbf{D}}_{1,\mathcal{J}_1}} \mathbf{D}^0_1\right] \right\} \dots \\
&\quad \mathbb{E}_{\mathcal{J}_k} \left\{ \operatorname{Tr}\left[\mathbf{D}^{0\top}_k (\mathbf{I}_{m_k} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}) \mathbf{D}^0_k\right] \right\} \dots \mathbb{E}_{\mathcal{J}_K} \left\{ \operatorname{Tr}\left[\mathbf{D}^{0\top}_K \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}} \Big\{ \operatorname{Tr}\Big[\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}} \Big] \Big\} \\
&= \frac{\mathbb{E}\{w^2\}}{2} \mathbb{E}_{\mathcal{J}} \Big\{ \sum_{k\in[K]} \operatorname{Tr}\left[\mathbf{P}_{\widetilde{\mathbf{D}}_{1,J_1}}\right] \dots \operatorname{Tr}\left[\mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right] \dots \operatorname{Tr}\left[\mathbf{P}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\right] \Big\} \\
&\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\} \dots \\
&\quad \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\} \dots \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\} \dots \\
&\quad \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\} \dots \mathbb{E}_{\mathcal{J}_K} \left\{ \operatorname{Tr}\left[\mathbf{H}_{\widetilde{\mathbf{D}}_{K,\mathcal{J}_K}}\right] \right\}.
\end{aligned} \tag{4.62}$$

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 $\operatorname{Tr}\left[\mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}}\right] = s_k - s_k = 0$. Adding the terms in (4.62), we obtain the expression in (4.25).

4.6.4 Proof of Lemma 4.5

Equation (4.26) follows from the definition of RIP and (4.27) follows from Gerschgorin's disk theorem [26, 84].

4.6.5 Proof of Lemma 4.8

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 (4.25) 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

$$\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[{\mathbf{D}^0_k}^\top \mathbf{P}_{\widetilde{\mathbf{D}}_{k,\mathcal{J}_k}} \mathbf{D}^0_k\right]\right\} = \mathbb{E}_{\mathcal{J}_k}\left\{\left\|\mathbf{P}_{\widetilde{\mathbf{D}}_{k,\mathcal{J}_k}} \mathbf{D}^0_{k,\mathcal{J}_k}\right\|_F^2\right\}. \tag{4.63}$$

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

$$\mathbb{E}_{\mathcal{J}_k}\left\{\left\|\mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}} \mathbf{D}^0_{k,\mathcal{J}_k}\right\|_F^2\right\} \overset{(a)}{=} \frac{s_k}{p_k}\left\|\mathbf{D}^0_k\right\|_F^2 = s_k, \tag{4.64}$$

where (a) follows from [24, Lemma 15]. If $\widetilde{\mathbf{D}}_k = \mathbf{D}_k$, then

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

where (b) is a direct consequence of Lemma 4.7; we can write $\mathbf{D}^0_k = \mathbf{D}_k\mathbf{C}_k^{-1} - \mathbf{V}_k\mathbf{T}_k$ where $\mathbf{C}_k = \mathrm{Diag}\left(\cos(\boldsymbol{\theta}_k)\right)$, $\mathbf{T}_k = \mathrm{Diag}\left(\tan(\boldsymbol{\theta}_k)\right)$ 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 [24, 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

$$\begin{aligned}\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[{\mathbf{D}^0_k}^\top(\mathbf{I}_{m_k} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}})\mathbf{D}^0_k\right]\right\} &= \mathbb{E}_{\mathcal{J}_k}\left\{\left\|(\mathbf{I}_{m_k} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}})\mathbf{D}^0_{k,\mathcal{J}_k}\right\|_F^2\right\} \\ &\overset{(f)}{\geq} \frac{s_k}{p_k}\|\boldsymbol{\theta}_k\|_2^2\left(1 - \frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k}\right),\end{aligned} \tag{4.65}$$

where (f) follows from similar arguments as in Gribonval et al. [24, Equation (72)]. Putting it all together, we have

$$\begin{aligned}\mathbb{E}\left\{\Delta\phi_1\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\} &\geq \frac{\mathbb{E}\{x^2\}}{2}\sum_{k\in[K]}\left(\prod_{\substack{i\in[K]\\ i\neq k}} s_i\right)\frac{s_k}{p_k}\|\boldsymbol{\theta}_k\|_2^2\left(1-\frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k}\right)\\ &= \frac{s\mathbb{E}\{x^2\}}{2}\sum_{k\in[K]}\frac{\|\boldsymbol{\theta}_k\|_2^2}{p_k}\left(1-\frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k}\right). \end{aligned} \tag{4.66}$$

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

$$\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[{\mathbf{D}^0}^+_{k,\mathcal{J}_k}\mathbf{D}^0_{k,\mathcal{J}_k}\right]\right\} = \mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{I}_{s_k}\right]\right\} = s_k, \tag{4.67}$$

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

$$\begin{aligned}\left|\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[{\mathbf{D}_{k,\mathcal{J}_k}}^+\mathbf{D}^0_k\right]\right\}\right| &\overset{(g)}{\leq} s_k\mathbb{E}_{\mathcal{J}_k}\left\{\left\|\mathbf{D}^+_{k,\mathcal{J}_k}\mathbf{D}^0_{k,\mathcal{J}_k}\right\|_2\right\}\\ &\leq s_k\mathbb{E}_{\mathcal{J}_k}\left\{\|\mathbf{D}^+_{k,\mathcal{J}_k}\|_2\|\mathbf{D}^0_{k,\mathcal{J}_k}\|_2\right\}\\ &\overset{(h)}{\leq} s_k\left(\frac{1}{\sqrt{1-\delta_{s_k}(\mathbf{D}_k)}}\right)\left(\sqrt{1+\delta_{s_k}(\mathbf{D}^0_k)}\right)\\ &\overset{(i)}{\leq} s_k\sqrt{\frac{1+\delta_k}{1-\delta_k}}, \end{aligned} \tag{4.68}$$

where (g) follows from the fact that for a square matrix $\mathbf{A}\in\mathbb{R}^{q\times q}$, $\mathrm{Tr}\left[\mathbf{A}\right]\leq q\|\mathbf{A}\|_2$, (h) follows from (4.26) and (4.28) and (i) follows from (4.32). Similar to [24, Equation (73)], we also have

$$\left|\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{I}_{s_k}-\mathbf{D}^+_{k,\mathcal{J}_k}\mathbf{D}^0_k\right]\right\}\right| \leq \frac{s_k}{p_k}\frac{\|\boldsymbol{\theta}_k\|_2^2}{2}+\frac{s_k^2}{p_k^2}\frac{A_kB_k}{1-\delta_k}\|\boldsymbol{\theta}_k\|_2. \tag{4.69}$$

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

$$\begin{aligned}
&\mathbb{E}\left\{\Delta\phi_4\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right\}\\
&\quad\geq -\lambda\mathbb{E}\{|x|\}\sum_{k\in[K]}\delta_{-k}\Bigg(\prod_{\substack{i\in[K]\\ i\neq k}} s_i\Bigg)\left(\frac{s_k}{p_k}\frac{\|\boldsymbol{\theta}_k\|_2^2}{2}+\frac{s_k^2}{p_k^2}\frac{A_kB_k}{1-\delta_k}\|\boldsymbol{\theta}_k\|_2\right)\\
&\quad= -\lambda s\mathbb{E}\{|x|\}\sum_{k\in[K]}\frac{\delta_{-k}}{p_k}\left(\frac{\|\boldsymbol{\theta}_k\|_2^2}{2}+\frac{s_k}{p_k}\frac{A_kB_k}{1-\delta_k}\|\boldsymbol{\theta}_k\|_2\right). \qquad (4.70)
\end{aligned}$$

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

$$\left|\mathbb{E}_{\mathcal{J}_k}\left\{\mathrm{Tr}\left[\mathbf{H}_{\widetilde{\mathbf{D}}_{k,\mathcal{J}_k}}\right]\right\}\right| \leq \mathbb{E}_{\mathcal{J}_k}\left\{s_k\left\|\mathbf{H}_{\widetilde{\mathbf{D}}_{k,\mathcal{J}_k}}\right\|_2\right\} \overset{(j)}{\leq} \frac{s_k}{1-\delta_k}, \qquad (4.71)$$

where (j) follows from (4.28) and (4.32). Similar to Gribonval et al. [24, Equation (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

$$\begin{aligned}
\mathbb{E}\{\Delta\phi_6\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\} &\geq -\frac{\lambda^2}{2}\sum_{k\in[K]}\Bigg(\prod_{\substack{i\in[K]\\ i\neq k}}\frac{s_i}{1-\delta_i}\Bigg)\left(\frac{s_k^2}{p_k^2}\frac{4A_kB_k}{(1-\delta_k)^2}\|\boldsymbol{\theta}_k\|_2\right)\\
&= -\frac{\lambda^2 s}{2}\sum_{k\in[K]}\frac{1}{p_k}\Bigg(\prod_{i\in[K]}\frac{1}{1-\delta_i}\Bigg)\left(\frac{s_k}{p_k}\frac{4A_kB_k}{1-\delta_k}\|\boldsymbol{\theta}_k\|_2\right). \qquad (4.72)
\end{aligned}$$

Adding (4.66), (4.70), and (4.72), we get (4.33).

4.6.6 Proof of Proposition 4.1

To show that $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right) > 0$, we use Lemma 4.8 and prove that the right hand side of (4.33) is positive under certain conditions. First, we ensure the conditions in (4.29) and (4.32) hold for Lemma 4.6 and Lemma 4.8, respectively. We set $\delta_k = \frac{1}{2}$,

$\delta_{s_k}(\mathbf{D}_k) = \frac{1}{2}$ and $\delta_{s_k}(\mathbf{D}_k^0) = \frac{1}{4}$, for $k \in [K]$. For $\varepsilon_k \leq 0.15$, this ensures:

$$\sqrt{1-\delta_{s_k}(\mathbf{D}_k)} \geq \sqrt{1-\delta_{s_k}(\mathbf{D}_k^0)} - \varepsilon_k, \text{ and } \max\left\{\delta_{s_k}(\mathbf{D}_k^0), \delta_{s_k}(\mathbf{D}_k)\right\} \leq \delta_k, \tag{4.73}$$

and implies $\delta_k < 1$ (condition for Lemmas 4.4 and 4.13). Next, we find conditions that guarantee:

$$\frac{s_k}{p_k}\frac{B_k^2}{1-\delta_k} + \bar{\lambda}\kappa_x^2\delta_{-k} \overset{(a)}{=} \frac{2B_k^2 s_k}{p_k} + \bar{\lambda}\kappa_x^2 (3)^{(K-1)/2} \leq \frac{1}{2}, \tag{4.74}$$

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 $\bar{\lambda} \leq \frac{1}{8 \times 3^{(K-1)/2}}$, given the fact that $\kappa_x^2 \leq 1$, (4.74) is satisfied.[11] Consequently, we can restate (4.33) as

$$\begin{aligned}&\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0 | \boldsymbol{\sigma}\right)\\ &\quad\geq \frac{s\mathbb{E}\{x^2\}}{4}\sum_{k\in[K]}\frac{\|\boldsymbol{\theta}_k\|_2}{p_k}\left[\|\boldsymbol{\theta}_k\|_2 - 8\left(3^{(K-1)/2} + 2^{(K+1)}\bar{\lambda}\right)\bar{\lambda}\kappa_x^2\frac{s_k}{p_k}A_k B_k\right].\end{aligned} \tag{4.75}$$

From [24, Proof of Proposition 2], we use the following relations:

$$B_k \leq B_k^0 + \varepsilon_k \leq B_k^0 + 1, \quad A_k \leq A_k^0 + 2B_k\varepsilon_k, \quad k \in [K], \tag{4.76}$$

where $A_k^0 \triangleq \left\|\mathbf{D}_k^{0\top}\mathbf{D}_k^0 - \mathbf{I}_{p_k}\right\|_F$ and $B_k^0 \triangleq \left\|\mathbf{D}_k^0\right\|_2$ and (4.76) follows from matrix norm inequalities [24]. Defining $\gamma_k \triangleq 16\left(3^{(K-1)/2} + 2^{(K+1)}\bar{\lambda}\right)\bar{\lambda}\kappa_x^2\frac{B_k^2 s_k}{p_k}$ for $k \in [K]$ and using $\kappa_x^2 \leq 1$, we have

$$\begin{aligned}\gamma_k &\leq 2\left(3^{(K-1)/2} + \frac{2^{(K+1)}}{8\times 3^{(K-1)/2}}\right)\left(\frac{1}{8\times 3^{(K-1)/2}}\right)\\ &\leq 2\left(\frac{1}{8} + \frac{4}{64}\right) \leq \frac{1}{2}.\end{aligned} \tag{4.77}$$

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

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

$$\begin{aligned}
\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0 | \boldsymbol{\sigma}\right) &\overset{(b)}{\geq} \frac{s\mathbb{E}\{x^2\}}{4} \sum_{k\in[K]} \frac{\varepsilon_k}{p_k} \left(\varepsilon_k - \frac{\gamma_k}{2}\frac{A_k}{B_k}\right) \\
&\overset{(c)}{\geq} \frac{s\mathbb{E}\{x^2\}}{4} \sum_{k\in[K]} \frac{\varepsilon_k}{p_k} \left(\varepsilon_k - \frac{\gamma_k}{2}\frac{A_k^0 + 2B_k\varepsilon_k}{B_k}\right) \\
&\geq \frac{s\mathbb{E}\{x^2\}}{4} \sum_{k\in[K]} \frac{\varepsilon_k}{p_k} \left(\varepsilon_k(1-\gamma_k) - \frac{\gamma_k}{2}\frac{A_k^0}{B_k}\right) \\
&\overset{(d)}{\geq} \frac{s\mathbb{E}\{x^2\}}{8} \sum_{k\in[K]} \frac{\varepsilon_k}{p_k} \left(\varepsilon_k - \gamma_k\frac{A_k^0}{B_k}\right), \qquad (4.78)
\end{aligned}$$

where (b) follows from (4.75), (c) follows from (4.76), and (d) follows from (4.77). Hence, we can write

$$\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}; \mathbf{D}_{1:K}^0 | \boldsymbol{\sigma}\right) \geq \frac{s\mathbb{E}\{x^2\}}{8} \sum_{k\in[K]} \frac{\varepsilon_k}{p_k} \left(\varepsilon_k - \varepsilon_{k,\min}(\bar{\lambda})\right), \qquad (4.79)$$

where we define

$$\begin{aligned}
\varepsilon_{k,\min}(\bar{\lambda}) &\triangleq \gamma_k \frac{A_k^0}{B_k} \\
&= 16\left(3^{(K-1)/2} + 2^{(K+1)}\bar{\lambda}\right)\bar{\lambda}\kappa_x^2 \frac{s_k}{p_k} A_k^0 B_k \\
&= \frac{2}{3^{(K+1)/2}}\left(3^{(K-1)/2} + 2^{(K+1)}\bar{\lambda}\right)\bar{\lambda} C_{k,\min}, \qquad (4.80)
\end{aligned}$$

and $C_{k,\min}$ is defined in (4.7). The lower bound in (4.79) 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 $\varepsilon_k \in (\varepsilon_{k,\min}(\bar{\lambda}), 0.15]$, $k \in [K]$.

4.6.7 Proof of Lemma 4.10

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

$$\begin{aligned}
\|\mathbf{D}_{\mathcal{J}}^\top \mathbf{d}_j\|_1 &\overset{(a)}{\leq} \|{\mathbf{D}_{\mathcal{J}}^0}^\top \mathbf{d}_j^0\|_1 + \|{\mathbf{D}_{\mathcal{J}}^0}^\top(\mathbf{d}_j - \mathbf{d}_j^0)\|_1 + \|(\mathbf{D}_{\mathcal{J}} - \mathbf{D}_{\mathcal{J}}^0)^\top \mathbf{d}_j\|_1 \\
&\leq \mu_s(\mathbf{D}^0) + \sqrt{s}\left[\|{\mathbf{D}_{\mathcal{J}}^0}^\top(\mathbf{d}_j - \mathbf{d}_j^0)\|_2 + \|(\mathbf{D}_{\mathcal{J}} - \mathbf{D}_{\mathcal{J}}^0)^\top \mathbf{d}_j\|_2\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mu_s(\mathbf{D}^0) + \sqrt{s}\Big[\Big\| \bigotimes {\mathbf{D}^0_{k,\mathcal{J}_k}}^\top \Big\|_2 \Big\| \bigotimes \left(\mathbf{d}_{k,j_k} - \mathbf{d}^0_{k,j_k}\right) \Big\|_2 \\
&\qquad + \Big\| \bigotimes \mathbf{D}_{k,\mathcal{J}_k} - \bigotimes \mathbf{D}^0_{k,\mathcal{J}_k} \Big\|_2 \|\mathbf{d}_j\|_2 \Big] \\
&\overset{(b)}{\leq} \mu_s(\mathbf{D}^0) + \sqrt{s}\Big[\Big(\prod_{k\in[K]} \sqrt{1+\delta_{s_k}(\mathbf{D}^0_k)} \Big) \\
&\qquad \Big(\sum_{k\in[K]} \Big\| \widetilde{\mathbf{d}}_{1,j_1} \Big\|_2 \dots \|\mathbf{d}_{k,j_k} - \mathbf{d}^0_{k,j_k}\|_2 \dots \Big\| \widetilde{\mathbf{d}}_{k,j_K} \Big\|_2 \Big) \\
&\quad + \sum_{k\in[K]} \Big\| \widetilde{\mathbf{D}}_{1,\mathcal{J}_1} \Big\|_2 \dots \|\mathbf{D}_{k,\mathcal{J}_k} - \mathbf{D}^0_{k,\mathcal{J}_k}\|_2 \dots \Big\| \widetilde{\mathbf{D}}_{k,\mathcal{J}_k} \Big\|_2 \Big] \\
&\overset{(c)}{\leq} \mu_s(\mathbf{D}^0) + \sqrt{s}\Big[\Big(\prod_{k\in[K]} \sqrt{1+\delta_{s_k}(\mathbf{D}^0_k)} \Big) \Big(\sum_{k\in[K]} \varepsilon_k \Big) \\
&\qquad + \sum_{k\in[K]} \Big(\prod_{\substack{i\in[K] \\ i\neq k}} \Big\| \widetilde{\mathbf{D}}_{i,\mathcal{J}_i} \Big\|_2 \Big) \varepsilon_k \Big] \\
&\overset{(d)}{\leq} \mu_s(\mathbf{D}^0) + 2(1.5)^{K/2}\sqrt{s}\Big(\sum_{k\in[K]} \varepsilon_k \Big), \qquad (4.81)
\end{aligned}
$$

where (a) follows from the triangle inequality, (b) follows from (4.20), (c) follows from (4.27), and, (d) follows from substituting the upper bound value from (4.38) for $\delta_{s_k}(\mathbf{D}^0_k)$. For $\widetilde{\mathbf{D}}_i = \mathbf{D}^0_i$, $\Big\| \mathbf{D}^0_{i,\mathcal{J}_i} \Big\|_2 \leq \sqrt{1+\delta_{s_i}(\mathbf{D}^0_i)} \leq \sqrt{\frac{5}{4}} < 1.5$ and for $\widetilde{\mathbf{D}}_i = \mathbf{D}_i$, according to (4.76), we have $\|\mathbf{D}_{i,\mathcal{J}_i}\|_2 \leq \Big\| \mathbf{D}^0_{i,\mathcal{J}_i} \Big\|_2 + \varepsilon_i \leq \sqrt{\frac{5}{4}} + 0.15 < 1.5$.

4.6.8 Proof of Proposition 4.2

We follow a similar approach to Gribonval et al. [24]. We show that the conditions in (4.37) hold for Lemma 4.9. We have

$$
\begin{aligned}
&\Big\| \mathbf{y} - \Big(\bigotimes \mathbf{D}_k \Big) \mathbf{x} \Big\|_2 \\
&\leq \Big\| \Big(\bigotimes \mathbf{D}^0_{k,\mathcal{J}_k} - \bigotimes \mathbf{D}_{k,\mathcal{J}_k} \Big) \mathbf{x}_{\mathcal{J}} \Big\|_2 + \|\mathbf{w}\|_2 \\
&\leq M_x \sum_{k\in[K]} \big\| \widetilde{\mathbf{D}}_{1,\mathcal{J}_1} \otimes \dots \otimes \left(\mathbf{D}^0_{k,\mathcal{J}_k} - \mathbf{D}_{k,\mathcal{J}_k} \right) \otimes \dots \otimes \widetilde{\mathbf{D}}_{K,\mathcal{J}_K} \big\|_2 + M_w \\
&\leq M_x \sum_{k\in[K]} \Big\| \widetilde{\mathbf{D}}_{1,\mathcal{J}_1} \Big\|_2 \dots \big\| \mathbf{D}^0_{k,\mathcal{J}_k} - \mathbf{D}_{k,\mathcal{J}_k} \big\|_2 \dots \Big\| \widetilde{\mathbf{D}}_{K,\mathcal{J}_K} \Big\|_2 + M_w
\end{aligned}
$$

$$\leq M_x \sum_{k\in[K]} \Bigg(\prod_{\substack{i\in[K]\\ i\neq k}} \left\| \widetilde{\mathbf{D}}_{i,\mathcal{J}_i} \right\|_2 \Bigg) \varepsilon_k + M_w$$

$$\overset{(a)}{\leq} (1.5)^{(K-1)/2} M_x \sum_{k\in[K]} \varepsilon_k + M_w, \tag{4.82}$$

where (a) follows from (4.34) 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 (4.76), 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

$$\begin{aligned}
&\lambda(1-2\mu_s(\mathbf{D})) - \left\| \mathbf{y} - \left(\bigotimes \mathbf{D}_k\right) \mathbf{x} \right\|_2 \\
&\geq \lambda(1-2\mu_s(\mathbf{D})) - (1.5)^{(K-1)/2} M_x \sum_{k\in[K]} \varepsilon_k - M_w \\
&\overset{(b)}{\geq} \lambda(1-2\mu_s(\mathbf{D}^0)) - (1.5)^{K/2} \left(4\lambda\sqrt{s} + (1.5)^{-1/2} M_x\right) \sum_{k\in[K]} \varepsilon_k - M_w \\
&\overset{(c)}{\geq} \lambda(1-2\mu_s(\mathbf{D}^0)) - 3(1.5)^{K/2} M_x \sum_{k\in[K]} \varepsilon_k - M_w \\
&= 3(1.5)^{K/2} M_x \Bigg(K\bar{\lambda} C_{\max} - \sum_{k\in[K]} \varepsilon_k \Bigg) - M_w,
\end{aligned} \tag{4.83}$$

where (b) follows from (4.39) and (c) follows from (4.37) ($2\lambda\sqrt{s} \leq x_{\min}\sqrt{s} \leq M_x$) and (4.39). If $\varepsilon_k < C_{\max}\bar{\lambda}$, $k \in [K]$, the assumption on the noise level in (4.36) implies that the right-hand side of (4.83) 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 4.9, $\widehat{\mathbf{x}}$ 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 $\Delta\phi_{\mathbb{P}}\left(\mathbf{D}_{1:K}, \mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right) = \Delta f_{\mathbb{P}}\left(\mathbf{D}_{1:K}, \mathbf{D}_{1:K}^0\right)$.

4.6.9 Proof of Lemma 4.12

According to Lemma 4.11, we have to upper bound $\mathbb{E}\left\{\sup_{\mathbf{D}_k\in\overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0),k\in[K]} \left|\frac{1}{N}\sum_{n\in[N]} \beta_n h_n(\mathbf{D}_{1:K})\right|\right\}$. 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 $C_{\mathbf{D}_{1:K}} = \sqrt{\frac{K}{N}} \sum_{k\in[K]} \left(L_k \sum_{i\in[m_k]} \sum_{j\in[p_k]} \zeta_{ij}^k (\mathbf{D}_k - \mathbf{D}_k^0)_{ij} \right)$,

where $\{\beta_n\}_{n=1}^N$'s and $\left\{\zeta_{ij}^k\right\}, k \in [K], i \in [m_k], j \in [p_k]$'s are independent standard

Gaussian vectors. We have

$$\begin{aligned}\mathbb{E}\left\{\left|A_{\mathbf{D}_{1:K}}-A_{\mathbf{D}'_{1:K}}\right|^2\right\} &= \frac{1}{N^2}\left|\sum_{n\in[N]} h_n(\mathbf{D}_{1:K})-h_n(\mathbf{D}'_{1:K})\right|^2\\ &\overset{(a)}{\leq} \frac{1}{N}\Big(\sum_{k\in[K]} L_k\|\mathbf{D}_k-\mathbf{D}'_k\|_F\Big)^2\\ &\overset{(b)}{\leq} \frac{K}{N}\sum_{k\in[K]} L_k^2\|\mathbf{D}_k-\mathbf{D}'_k\|_F^2\\ &= \mathbb{E}\left\{\left|C_{\mathbf{D}_{1:K}}-C_{\mathbf{D}'_{1:K}}\right|^2\right\}, \end{aligned} \tag{4.84}$$

where (a) follows from coordinate-wise Lipschitz continuity of h and (b) follows from Cauchy-Schwartz inequality. Hence, using Slepian's Lemma [61], we get

$$\begin{aligned}\mathbb{E}\Big\{\sup_{\substack{\mathbf{D}_k\in\overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0)\\ k\in[K]}} A_{\mathbf{D}_{1:K}}\Big\} &\leq \mathbb{E}\Big\{\sup_{\substack{\mathbf{D}_k\in\overline{\mathcal{B}}_{\varepsilon_k}(\mathbf{D}_k^0)\\ k\in[K]}} C_{\mathbf{D}_{1:K}}\Big\}\\ &= \sqrt{\frac{K}{N}}\Big(\sum_{k\in[K]} L_k\varepsilon_k\mathbb{E}\big\{\|\boldsymbol{\zeta}^k\|_F\big\}\Big)\\ &= \sqrt{\frac{K}{N}}\Big(\sum_{k\in[K]} L_k\varepsilon_k\sqrt{m_k p_k}\Big). \end{aligned} \tag{4.85}$$

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

4.6.10 Proof of Lemma 4.14

We expand $\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)$ according to (4.59) and bound each term of the sum separately. Looking at the first term, we get

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

$$\overset{(b)}{\leq} \frac{1}{2}\left\|\mathbf{x}\right\|_2^2 \Bigg(\prod_{k\in[K]} \left\|\mathbf{D}^0_{k,\mathcal{J}_k}\right\|_2^2 \Bigg) \Bigg(\sum_{k\in[K]} \left\| \mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}} \right\|_2 \Bigg(\prod_{\substack{i\in[K]\\ i\neq k}} \left\| \mathbf{P}_{\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}} \right\|_2 \Bigg) \Bigg)$$

$$\overset{(c)}{\leq} M_x^2 \Bigg(\prod_{k\in[K]} \left(1+\delta_{s_k}(\mathbf{D}^0_k)\right) \Bigg) \Bigg(\sum_{k\in[K]} (1-\delta_k)^{-1/2} \|\mathbf{D}_k - \mathbf{D}^0_k\|_F \Bigg), \tag{4.86}$$

where (a) follows from (4.60), (b) follows from the fact that $\left\|\mathbf{D}^0_{\mathcal{J}}\right\|_2 = \prod_{k\in[K]} \left\|\mathbf{D}^0_{k,\mathcal{J}_k}\right\|_2$, and (c) follows from the definition of RIP, equation (4.48), 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}^0_{1:K}|\boldsymbol{\sigma}\right)\right| \\
&\leq \|\mathbf{w}\|_2 \|\mathbf{x}\|_2 \Bigg(\prod_{k\in[K]} \left\|\mathbf{D}^0_{k,\mathcal{J}_k}\right\|_2^2 \Bigg) \Bigg(\sum_{k\in[K]} \left\| \mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}} \right\|_2 \Bigg(\prod_{\substack{i\in[K]\\ i\neq k}} \left\| \mathbf{P}_{\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}} \right\|_2 \Bigg) \Bigg) \\
&\overset{(d)}{\leq} 2M_w M_x \Bigg(\prod_{k\in[K]} \left(1+\delta_{s_k}(\mathbf{D}^0_k)\right)^{1/2} \Bigg) \Bigg(\sum_{k\in[K]} (1-\delta_k)^{-1/2} \|\mathbf{D}_k - \mathbf{D}^0_k\|_F \Bigg), \\
&\left|\Delta\phi_3\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right| \\
&\leq \frac{1}{2}\|\mathbf{w}\|_2^2 \Bigg(\sum_{k\in[K]} \left\| \mathbf{P}_{\mathbf{D}^0_{k,\mathcal{J}_k}} - \mathbf{P}_{\mathbf{D}_{k,\mathcal{J}_k}} \right\|_2 \Bigg(\prod_{\substack{i\in[K]\\ i\neq k}} \left\| \mathbf{P}_{\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}} \right\|_2 \Bigg) \Bigg) \\
&\leq M_w^2 \Bigg(\sum_{k\in[K]} (1-\delta_k)^{-1/2} \|\mathbf{D}_k - \mathbf{D}^0_k\|_F \Bigg), \\
&\left|\Delta\phi_4\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right| \\
&= \lambda \|\boldsymbol{\sigma}_{\mathcal{J}}\|_2 \|\mathbf{x}\|_2 \Bigg(\prod_{k\in[K]} \left\|\mathbf{D}^0_{\mathcal{J}_k}\right\|_2 \Bigg) \Bigg(\sum_{k\in[K]} \left\| {\mathbf{D}^0}^+_{k,\mathcal{J}_k} - \mathbf{D}^+_{k,\mathcal{J}_k} \right\|_2 \Bigg(\prod_{\substack{i\in[K]\\ i\neq k}} \left\| \widetilde{\mathbf{D}}^+_{i,\mathcal{J}_i} \right\|_2 \Bigg) \Bigg) \\
&\overset{(e)}{\leq} 2\lambda\sqrt{s} M_x \Bigg(\prod_{k\in[K]} \left(1+\delta_{s_k}(\mathbf{D}^0_k)\right)^{1/2} \Bigg) \\
&\qquad \Bigg(\sum_{k\in[K]} (1-\delta_k)^{-1} \Bigg(\prod_{\substack{i\in[K]\\ i\neq k}} (1-\delta_i)^{-1/2} \Bigg) \|\mathbf{D}_k - \mathbf{D}^0_k\|_F \Bigg), \\
&\left|\Delta\phi_5\left(\mathbf{D}_{1:K};\mathbf{D}^0_{1:K}|\boldsymbol{\sigma}\right)\right| \\
&= \lambda \|\boldsymbol{\sigma}_{\mathcal{J}}\|_2 \|\mathbf{w}\|_2 \Bigg(\sum_{k\in[K]} \left\| {\mathbf{D}^0}^+_{k,\mathcal{J}_k} - \mathbf{D}^+_{k,\mathcal{J}_k} \right\|_2 \Bigg(\prod_{\substack{i\in[K]\\ i\neq k}} \left\| \widetilde{\mathbf{D}}^+_{i,\mathcal{J}_i} \right\|_2 \Bigg) \Bigg)
\end{aligned}$$

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

$$\begin{aligned}
&\left|\Delta\phi_6\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\right| \\
&= \frac{\lambda^2}{2}\left\|\boldsymbol{\sigma}_{\mathcal{J}}\right\|_2^2\Big(\sum_{k\in[K]}\Big\|\mathbf{H}_{\mathbf{D}^0_{k,\mathcal{J}_k}}-\mathbf{H}_{\mathbf{D}_{k,\mathcal{J}_k}}\Big\|_2\Big(\prod_{\substack{i\in[K]\\ i\neq k}}\Big\|\mathbf{H}_{\widetilde{\mathbf{D}}_{i,\mathcal{J}_i}}\Big\|_2\Big)\Big) \\
&\overset{(f)}{\leq} \lambda^2 s\Big(\sum_{k\in[K]}(1-\delta_k)^{-\frac{3}{2}}\Big(\prod_{\substack{i\in[K]\\ i\neq k}}(1-\delta_i)^{-1}\Big)\|\mathbf{D}_k-\mathbf{D}_k^0\|_F\Big),
\end{aligned} \tag{4.87}$$

where (e) and (f) follow from (4.28) and (4.48). Adding all the terms together, we get

$$\left|\Delta\phi_{\mathbf{y}}\left(\mathbf{D}_{1:K};\mathbf{D}_{1:K}^0|\boldsymbol{\sigma}\right)\right| \leq \sum_{k\in[K]} L_k\|\mathbf{D}_k-\mathbf{D}_k^0\|_F. \tag{4.88}$$

where L_k is defined in (4.49).

4.6.11 Proof of the coherence relation for KS dictionaries

To prove (1.4), 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

$$\begin{aligned}
\mu_s(\mathbf{X}) &= \max_{|\mathcal{J}|\leq s}\max_{j\notin\mathcal{J}}\|\mathbf{X}_{\mathcal{J}}^\top\mathbf{x}_j\|_1 \\
&= \max_{\substack{|\mathcal{J}_k|\leq s_k\\ k\in[K]}}\max_{\mathcal{A}}\Big\|\Big(\bigotimes\mathbf{X}_{k,\mathcal{J}_k}^\top\Big)\Big(\bigotimes\mathbf{x}_{k,j_k}\Big)\Big\|_1 \\
&= \max_{\substack{|\mathcal{J}_k|\leq s_k\\ k\in[K]}}\max_{\mathcal{A}}\Big\|\bigotimes\mathbf{X}_{k,\mathcal{J}_k}^\top\mathbf{x}_{k,j_k}\Big\|_1 \\
&= \max_{\substack{|\mathcal{J}_k|\leq s_k\\ k\in[K]}}\max_{\mathcal{A}}\prod_{k\in[K]}\Big\|\mathbf{X}_{k,\mathcal{J}_k}^\top\mathbf{x}_{k,j_k}\Big\|_1 \\
&\leq \max_{k\in[K]}\mu_{s_k}(\mathbf{X}_k)\Big(\prod_{\substack{i\in[K],\\ i\neq k}}(1+\mu_{s_i-1}(\mathbf{X}_i))\Big).
\end{aligned} \tag{4.89}$$

Chapter 5

Learning Mixtures of Separable Dictionaries for Tensor Data

This chapter addresses the problem of learning sparse representations of tensor data using a mixture of separable dictionaries. This model better captures the structure of tensor data by generalizing the Kronecker-structured dictionary learning model. Various algorithms are developed to solve the problem of learning mixture of separable dictionaries in both batch and online settings. Numerical experiments are provided to show the usefulness of the proposed model and the efficacy of the developed algorithms for synthetic data representation and real-world data image denoising.[1]

5.1 Introduction

In Chapters 3 and 4, we focused on the Kronecker-structured dictionary learning (KS-DL) model for tensor data representation. While existing KS-DL methods enjoy lower sample/computational complexity and better storage efficiency over unstructured DL [12], the KS-DL model makes a strong separability assumption among different modes of tensor data, which is often too restrictive for many classes of data [13]. This results in an unfavorable tradeoff between model compactness and representation power. In this chapter, we overcome this limitation by taking advantage a generalization of KS-DL that we interchangeably refer to as *learning a mixture of separable dictionaries* or *low separation-rank DL* (LSR-DL). More specifically, the separation rank of a matrix

[1]The work presented in this chapter was done in collaboration with graduate student Mohsen Ghassemi and the results have been published in the Proceedings of 2017 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing [62] and submitted to IEEE Transactions on Signal Processing [88]. This work includes theoretical and experimental contributions. In this chapter, we focus solely on algorithms and numerical experiments.

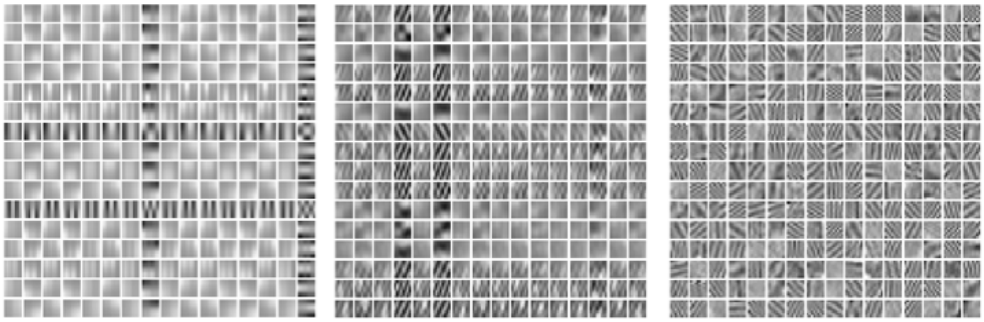


Figure 5.1: Dictionary atoms for representing RGB image `Barbara` for separation rank (left-to-right) 1, 4, and 256.

$\mathbf{A}$ is defined as the minimum number of KS matrices whose sum equals $\mathbf{A}$ [89, 90]. The LSR-DL model interpolates between the under-parameterized KS-DL model (a special case of LSR-DL model with separation rank 1) and the over-parameterized unstructured model. Figure 5.1 provides an illustrative example of the usefulness of LSR-DL for image data representation. While KS-DL learns dictionary atoms that cannot reconstruct diagonal structures because of horizontal/vertical (DCT-like) structures, increasing the separation rank results in dictionary atoms with pronounced diagonal structures.

5.1.1 Main Contributions

We develop DL algorithms that enforce the LSR structure on the underlying dictionary. In this regard, we first study a simple block coordinate descent-based algorithm called SubDil that alternates between solving the problem with respect to each coordinate dictionary. Despite its good performance in learning structured dictionaries in certain settings, this algorithm suffers in accuracy and speed when searching for higher separation-rank dictionaries. To address this issue, we take advantage of a connection between LSR matrices and low-rank tensors [62] which allows us to leverage ideas and tools from the tensor recovery literature. We provide a LSR-DL algorithm called STARK that takes advantage of a convex regularizer to impose LSR structure on the underlying dictionary. However, this algorithm only outputs the dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ and not the coordinate dictionaries $\{\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}\}_{k=1}^{K}$. Moreover, this method does not allow for explicit tuning of the separation rank to control the number

of parameters of the model. We then develop a LSR-DL algorithm called TeFDiL that employs tensor CP decomposition to impose LSR structure on the dictionary. This algorithm estimates the coordinate dictionaries and allows explicit tuning of the separation rank. All our provided LSR-DL algorithms allow tuning of the number of KS components in the dictionary to avoid underfitting (inadequately small model) and overfitting (excessively large model). We also provide a variation of SubDil called OSubDil for online data representation.

Finally, we validate the usefulness of LSR-DL using numerical experiments on synthetic data representation and real-world image data denoising and provide a comparison of the performance of our algorithms with unstructured and KS DL.

5.1.2 Relation to Prior Work

In terms of computational algorithms, several works have proposed methods for learning KS dictionaries [14–16, 69, 91]. Focusing on LSR-DL, Dantas et al. have proposed an LSR-DL algorithm that employs a convex regularizer to impose LSR structure on the dictionary for tensors of order $K = 2$ [70]. In contrast, our regularization-based algorithm STARK can find LSR dictionaries for tensors of any order $K \geq 2$. Moreover, an algorithm based on dictionary rearrangement that uses CP decomposition to perform LSR-DL has been proposed in [92]. The dictionary update stage of this method is a projected gradient descent algorithm that involves a CP decomposition after every gradient step. Our TeFDiL algorithm is also based on dictionary rearrangement and uses CP decomposition, but only requires a single CP decomposition at the end of each dictionary update stage. Finally, while there exist a number of online algorithms for DL [54, 93, 94], the online algorithm developed here is the first that enables learning of structured (either KS or LSR) dictionaries.

5.2 Problem Formulation

We propose the LSR-DL model in which the separation rank of the underlying dictionary is relatively small so that $1 \leq \mathfrak{R}(\mathbf{D}^0) \ll \min\{m, p\}$, where $\mathfrak{R}(\mathbf{D}^0)$ denotes the

separation rank of $\mathbf{D}^0$. This generalizes the KS-DL model to a generating dictionary of the form

$$\mathbf{D}^0 = \sum_{r=1}^{R} [\mathbf{D}_K^r]^0 \otimes [\mathbf{D}_{K-1}^r]^0 \otimes \cdots \otimes [\mathbf{D}_1^r]^0, \tag{5.1}$$

where R is the separation rank of $\mathbf{D}^0$. Note that the KS-DL model corresponds to separation rank 1.

Consequently, defining $\mathcal{D}_{KS}^R \triangleq \{\mathbf{D} \in \mathcal{D} | \mathfrak{R}(\mathbf{D}) \leq R\}$, the empirical rank-constrained LSR-DL problem is

$$\min_{\mathbf{D} \in \mathcal{D}_{KS}^R} \frac{1}{N} \sum_{n=1}^{N} f_{\mathbf{y}_n}(\mathbf{D}), \quad \text{where} \quad f_{\mathbf{y}_n}(\mathbf{D}) \triangleq \inf_{\mathbf{x} \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\mathbf{y}_n - \mathbf{D}\mathbf{x}_n\|_2^2 + \lambda \|\mathbf{x}_n\|_1 \right\}. \tag{5.2}$$

Lemma 5.1. *Any Kth-order KS matrix $\mathbf{A} = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_K$ can be rearranged as a rank-1, Kth-order tensor $\underline{\mathbf{A}}^\pi = \mathbf{a}_K \circ \cdots \circ \mathbf{a}_2 \circ \mathbf{a}_1$ with $\mathbf{a}_k \triangleq \mathrm{vec}(\mathbf{A}_k)$.*

It follows immediately from Lemma 5.1 that if $\mathbf{D} = \sum_{r=1}^{R} \mathbf{D}_1^r \otimes \mathbf{D}_2^r \otimes \cdots \otimes \mathbf{D}_K^r$, then we can rearrange matrix $\mathbf{D}$ into the tensor $\underline{\mathbf{D}}^\pi = \sum_{r=1}^{R} \mathbf{d}_K^r \circ \mathbf{d}_{K-1}^r \circ \cdots \circ \mathbf{d}_1^r$, where $\mathbf{d}_k^r = \mathrm{vec}(\mathbf{D}_k^r)$. Therefore, we have the following equivalence:

$$\mathfrak{R}(\mathbf{D}) \leq R \Longleftrightarrow \mathrm{rank}(\underline{\mathbf{D}}^\pi) \leq R.$$

This correspondence between separation rank and tensor rank highlights a challenge with the LSR-DL problem: finding the rank of a tensor is NP-hard [95] and thus so is finding the separation rank of a matrix. This makes Problem (5.2) in its current form (and its variants) intractable. To overcome this, we introduce two tractable relaxations to the rank-constrained Problem (5.2) that do not require explicit computation of the tensor rank.

The first relaxation uses a convex regularization term to implicitly impose low tensor rank structure on $\underline{\mathbf{D}}^\pi$, which results in a low separation rank $\mathbf{D}$. The resulting empirical

regularization-based LSR-DL problem is

$$\min_{\mathbf{D}\in\mathcal{D}} \frac{1}{N}\sum_{n=1}^{N} f_{\mathbf{y}_n}(\mathbf{D}) + \lambda_1 g_1(\underline{\mathbf{D}}^{\pi}), \tag{5.3}$$

where $g_1(\underline{\mathbf{D}}^{\pi})$ is a convex regularizer to enforce low-rank structure on $\underline{\mathbf{D}}^{\pi}$. The second relaxation is a *factorization-based LSR-DL formulation* in which the LSR dictionary is explicitly written in terms of its coordinate dictionaries. The resulting empirical risk minimization problem is

$$\min_{\{\mathbf{D}_k^r\}:\ \sum_{r=1}^{R}\bigotimes_{k=1}^{K}\mathbf{D}_k^r\in\mathcal{D}} \frac{1}{N}\sum_{n=1}^{N}\inf_{\mathbf{x}\in\mathbb{R}^p}\left\|\mathbf{y}-\Big(\sum_{r=1}^{R}\bigotimes_{k=1}^{K}\mathbf{D}_k^r\Big)\mathbf{x}\right\|^2 + \lambda\left\|\mathbf{x}\right\|_1. \tag{5.4}$$

Next, we propose algorithms to find solutions to Problems (5.3) and (5.4).

5.3 LSR-DL Algorithms

Solving Problems (5.3) and (5.4) require minimization with respect to (w.r.t.) $\mathbf{X}$. Therefore, similar to conventional DL algorithms, we introduce alternating minimization-type algorithms that at every iteration, first perform minimization of the objective function w.r.t. $\mathbf{X}$ (sparse coding stage) and then minimize the objective w.r.t. the dictionary (dictionary update stage).

5.3.1 STARK: A Regularization-based LSR-DL Algorithm

We first discuss an algorithm, which we term *STructured dictionAry learning via Regularized low-ranK Tensor Recovery (STARK)*, that helps solve the regularized LSR-DL problem given in (5.3) using the Alternating Direction Method of Multipliers (ADMM) [96].

The regularizer that we use here is a commonly used convex proxy for the tensor rank function, the *sum-trace-norm* [97], which is defined as the average of the trace (nuclear) norms of the unfoldings of the tensor: $g_1(\underline{\mathbf{D}}^{\pi}) = \|\underline{\mathbf{D}}^{\pi}\|_{\text{str}} \triangleq \sum_{k=1}^{K}\left\|\underline{\mathbf{D}}^{\pi}_{(k)}\right\|_{\text{tr}}$.

The main novelty in solving (5.3) with $g_1(\underline{\mathbf{D}}^{\pi}) = \|\underline{\mathbf{D}}^{\pi}\|_{\text{str}}$ is the dictionary update stage. This stage, which involves updating $\mathbf{D}$ for a fixed set of sparse codes $\mathbf{X}$, is

particularly challenging for gradient-based methods because the dictionary update involves interdependent nuclear norms of different unfoldings of the rearranged tensor $\underline{\mathbf{D}}^{\pi}$. Inspired by many works in the literature on low-rank tensor estimation [97–99], we instead suggest the following reformulation of the dictionary update stage of (5.3):

$$\min_{\mathbf{D}\in\mathcal{D},\underline{\mathbf{W}}_1,\cdots,\underline{\mathbf{W}}_K} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \lambda_1 \sum_{k=1}^{K} \left\|\mathbf{W}_{k,(k)}\right\|_{\mathrm{tr}}, \quad \text{s.t.} \quad \forall k \quad \underline{\mathbf{W}}_k = \underline{\mathbf{D}}^{\pi}, \tag{5.5}$$

where $\mathbf{W}_{k,(k)}$ denotes the kth-mode unfolding of $\underline{\mathbf{W}}_k$. In this formulation, although the nuclear norms depend on one another through the introduced constraint, we can decouple the minimization problem into separate subproblems. To solve this problem, we first find a solution to the problem without the constraint $\mathbf{D}\in\mathcal{D}$, then project the solution onto $\mathcal{D}$ by normalizing the columns of $\mathbf{D}$. We can solve the objective function (5.5) (without $\mathbf{D}\in\mathcal{D}$ contraint) using ADMM, which involves decoupling the problem into independent subproblems by forming the following augmented Lagrangian:

$$\mathcal{L}_\gamma = \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \sum_{k=1}^{K}\Big(\lambda_1\left\|\mathbf{W}_{k,(k)}\right\|_{\mathrm{tr}} - \langle\underline{\mathbf{A}}_k,\ \underline{\mathbf{D}}^{\pi}-\underline{\mathbf{W}}_k\rangle + \frac{\gamma}{2}\|\ \underline{\mathbf{D}}^{\pi}-\underline{\mathbf{W}}_k\|_F^2\Big), \tag{5.6}$$

where $\mathcal{L}_\gamma$ is shorthand for $\mathcal{L}_\gamma(\underline{\mathbf{D}}^{\pi},\{\underline{\mathbf{W}}_k\},\{\underline{\mathbf{A}}_k\})$. In order to find the gradient of (5.6) with respect to $\underline{\mathbf{D}}^{\pi}$, we rewrite the Lagrangian function in the following form:

$$\mathcal{L}_\gamma = \frac{1}{2}\|\widetilde{\mathbf{y}}-\mathcal{T}(\underline{\mathbf{D}}^{\pi})\|_2^2 + \sum_{k=1}^{K}\Big(\lambda_1\left\|\mathbf{W}_{k,(k)}\right\|_{\mathrm{tr}} - \langle\underline{\mathbf{A}}_k,\ \underline{\mathbf{D}}^{\pi}-\underline{\mathbf{W}}_k\rangle + \frac{\gamma}{2}\|\ \underline{\mathbf{D}}^{\pi}-\underline{\mathbf{W}}_k\|_F^2\Big), \tag{5.7}$$

where $\widetilde{\mathbf{y}} \triangleq \mathrm{vec}(\mathbf{Y})$ and the linear operator $\mathcal{T}(\underline{\mathbf{D}}^{\pi}) \triangleq \mathrm{vec}(\mathbf{D}\mathbf{X}) = \widetilde{\mathbf{X}}^{\top}\mathbf{\Pi}^{\top}\mathrm{vec}(\underline{\mathbf{D}}^{\pi})$, where $\widetilde{\mathbf{X}} = \mathbf{X}\otimes\mathbf{I}_m$ and $\mathbf{\Pi}$ is a permutation matrix such that $\mathrm{vec}(\underline{\mathbf{D}}^{\pi}) = \mathbf{\Pi}\,\mathrm{vec}(\mathbf{D})$. The procedure to find $\mathbf{\Pi}$ is explained in the Appendix. In the rest of this section, we discuss derivation of the update steps of ADMM.

ADMM Update Rules: Each iteration τ of ADMM consists of the following steps

[96]:

$$\underline{\mathbf{D}}^{\pi}(\tau) = \arg\min_{\underline{\mathbf{D}}^{\pi}} \mathcal{L}_{\gamma}(\underline{\mathbf{D}}^{\pi}, \underline{\mathbf{W}}_k(\tau-1), \underline{\mathbf{A}}_k(\tau-1)), \tag{5.8}$$

$$\underline{\mathbf{W}}_k(\tau) = \arg\min_{\underline{\mathbf{W}}_k} \mathcal{L}_{\gamma}(\underline{\mathbf{D}}^{\pi}(\tau), \underline{\mathbf{W}}_k, \underline{\mathbf{A}}_k(\tau-1)),\ \forall k \in [K], \tag{5.9}$$

$$\underline{\mathbf{A}}_k(\tau) = \underline{\mathbf{A}}_k(\tau-1) - \gamma\left(\underline{\mathbf{D}}^{\pi}(\tau) - \underline{\mathbf{W}}_k(\tau)\right),\ \forall\ k \in [K]. \tag{5.10}$$

The solution to (5.8) can be obtained by taking the gradient of $\mathcal{L}_{\gamma}(\cdot)$ w.r.t. $\underline{\mathbf{D}}^{\pi}$ and setting it to zero. Suppressing the iteration index τ for ease of notation, we have

$$\frac{\partial \mathcal{L}_{\gamma}}{\partial \underline{\mathbf{D}}^{\pi}} = \mathcal{T}^{*}(\mathcal{T}(\underline{\mathbf{D}}^{\pi}) - \widetilde{\mathbf{y}}) - \sum_{k=1}^{K} \underline{\mathbf{A}}_k + \sum_{k=1}^{K} \gamma\left(\underline{\mathbf{D}}^{\pi} - \underline{\mathbf{W}}_k\right), \tag{5.11}$$

where $\mathcal{T}^{*}(\mathbf{v}) = \mathrm{vec}^{-1}\left(\mathbf{\Pi}\widetilde{\mathbf{X}}\mathbf{v}\right)$ is the *adjoint* of the linear operator $\mathcal{T}$[99]. Setting the gradient to zero results in

$$\mathcal{T}^{*}(\mathcal{T}(\underline{\mathbf{D}}^{\pi})) + \gamma K\ \underline{\mathbf{D}}^{\pi} = \mathcal{T}^{*}(\widetilde{\mathbf{y}}) + \sum_{k=1}^{K} \left(\underline{\mathbf{A}}_k + \gamma \underline{\mathbf{W}}_k\right). \tag{5.12}$$

Equivalently, we have

$$\mathrm{vec}^{-1}\left(\left[\mathbf{\Pi}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\top}\mathbf{\Pi}^{\top} + \gamma K \mathbf{I}\right] \mathrm{vec}(\underline{\mathbf{D}}^{\pi})\right) = \mathrm{vec}^{-1}(\mathbf{\Pi}\widetilde{\mathbf{X}}\widetilde{\mathbf{y}}) + \sum_{k=1}^{K} \left(\underline{\mathbf{A}}_k + \gamma \underline{\mathbf{W}}_k\right). \tag{5.13}$$

Therefore, the update rule for $\underline{\mathbf{D}}^{\pi}$ is

$$\begin{aligned}\underline{\mathbf{D}}^{\pi}(\tau) = \mathrm{vec}^{-1}\Big(\Big[\mathbf{\Pi}^{T}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\top}\mathbf{\Pi} + \gamma K \mathbf{I}_{mp}\Big]^{-1} \\ \cdot \Big[\mathbf{\Pi}^{T}\widetilde{\mathbf{X}}\widetilde{\mathbf{y}} + \mathrm{vec}\Big(\sum_{k=1}^{K} \left(\underline{\mathbf{A}}_k(\tau-1) + \gamma \underline{\mathbf{W}}_k(\tau-1)\right)\Big)\Big]\Big).\end{aligned} \tag{5.14}$$

To update $\{\underline{\mathbf{W}}_k\}$, we can further split (5.9) into N independent subproblems (suppressing the index τ):

$$\min_{\underline{\mathbf{W}}_k}\ \mathcal{L}_{\mathcal{W}} = \lambda_1 \left\|\mathbf{W}_{k,(k)}\right\|_{\mathrm{tr}} - \langle \underline{\mathbf{A}}_k,\ \underline{\mathbf{D}}^{\pi} - \underline{\mathbf{W}}_k\rangle + \frac{\gamma}{2} \| \underline{\mathbf{D}}^{\pi} - \underline{\mathbf{W}}_k\|_F^2.$$

We can reformulate $\mathcal{L}_{\mathcal{W}}$ as

$$\lambda_1 \left\|\mathbf{W}_{k,(k)}\right\|_{\text{tr}} + \frac{\gamma}{2}\left\|\mathbf{W}_{k,(k)} - \left(\underline{\mathbf{D}}^{\pi}_{(k)} - \frac{\mathbf{A}_{k,(k)}}{\gamma}\right)\right\|_F^2 + \text{const.} \tag{5.15}$$

The minimizer of $\mathcal{L}_{\mathcal{W}}$ with respect to $\mathbf{W}_{k,(k)}$ is $\text{shrink}\left(\mathbf{D}^{\pi}_{(k)} - \frac{1}{\gamma}\mathbf{A}_{k,(k)},\ \frac{\lambda_1}{\gamma}\right)$ where $\text{shrink}(\mathbf{A}, z)$ applies soft thresholding at level z on the singular values of matrix $\mathbf{A}$ [100]. Therefore,

$$\underline{\mathbf{W}}_k(\tau) = \text{refold}\Big(\text{shrink}\left(\mathbf{D}^{\pi}_{(k)}(\tau) - \frac{1}{\gamma}\mathbf{A}_{k,(k)}(\tau-1),\ \frac{\lambda_1}{\gamma}\right)\Big), \tag{5.16}$$

where $\text{refold}(\cdot)$ is the inverse of the unfolding operator. Algorithm 1 summarizes this discussion and provides pseudocode for the dictionary update stage in STARK.

Algorithm 1 Dictionary Update in STARK for LSR-DL

Require: $\mathbf{Y}$, $\mathbf{\Pi}$, $\lambda_1 > 0$, $\gamma > 0$, $\mathbf{X}(t)$[2]

repeat
  Update $\underline{\mathbf{D}}^{\pi}$ according to update rule (5.14)
  for $k \in [K]$ **do**
    Update $\underline{\mathbf{W}}_k$ according to (5.16)
  end for
  for $k \in [K]$ **do**
    $\underline{\mathbf{A}}_k \leftarrow \underline{\mathbf{A}}_k - \gamma\left(\underline{\mathbf{D}}^{\pi} - \underline{\mathbf{W}}_k\right)$
  end for
until convergence
Normalize columns of $\mathbf{D}$
return $\mathbf{D}(t+1)$

5.3.2 TeFDiL: A Factorization-based LSR-DL Algorithm

While our experiments in Section 5.4 validate good performance of STARK, the algorithm finds the dictionary $\mathbf{D}$ and not the coordinate dictionaries $\{\mathbf{D}^r_k\}, k \in [K], r \in [R]$. Moreover, STARK only allows indirect control over the separation rank of the dictionary through the regularization parameter λ_1. This motivates developing a factorization-based LSR-DL algorithm that can find the coordinate dictionaries and allows for direct tuning of the separation rank to control the number of parameters of the model. To this

[2]In the body of Algorithms 1–3 we drop the iteration index t for simplicity.

end, we propose a factorization-based LSR-DL algorithm termed *Tensor Factorization-Based DL (TeFDiL)* in this section for solving Problem (5.4).

We discussed earlier in Section 5.3.1 that the error term $\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2$ can be reformulated as $\|\widetilde{\mathbf{y}} - \mathcal{T}(\underline{\mathbf{D}}^\pi)\|^2$ where $\mathcal{T}(\underline{\mathbf{D}}^\pi) = \widetilde{\mathbf{X}}^\top \mathbf{\Pi}^\top \operatorname{vec}(\underline{\mathbf{D}}^\pi)$. Thus, the dictionary update objective in (5.4) can be reformulated as $\|\widetilde{\mathbf{y}} - \mathcal{T}(\sum_{r=1}^R \mathbf{d}_K^r \circ \cdots \circ \mathbf{d}_1^r)\|^2$ where $\mathbf{d}_k^r = \operatorname{vec}(\mathbf{D}_k^r)$. To avoid the complexity of solving this problem, we resort to first obtaining an inexact solution by minimizing $\|\widetilde{\mathbf{y}} - \mathcal{T}(\underline{\mathbf{A}})\|^2$ over $\underline{\mathbf{A}}$ and then enforcing the low-rank structure by finding the rank-R approximation of the minimizer of $\|\widetilde{\mathbf{y}} - \mathcal{T}(\underline{\mathbf{A}})\|^2$. TeFDiL employs CP decomposition to find this approximation and thus enforce LSR structure on the updated dictionary.

Assuming the matrix of sparse codes $\mathbf{X}$ is full row-rank[3], then $\widetilde{\mathbf{X}}^\top$ is full column-rank and $\underline{\mathbf{A}} = \mathcal{T}^+(\widetilde{\mathbf{y}}) = \operatorname{vec}^{-1}\left(\mathbf{\Pi}\big(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\big)^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{y}}\right)$ minimizes $\|\widetilde{\mathbf{y}} - \mathcal{T}(\underline{\mathbf{A}})\|^2$. Now, it remains to solve the following problem to update $\{\mathbf{d}_k^r\}$:

$$\min_{\{\mathbf{d}_k^r\}} \left\| \sum_{r=1}^R \mathbf{d}_K^r \circ \cdots \circ \mathbf{d}_1^r - \mathcal{T}^+(\widetilde{\mathbf{y}}) \right\|_F^2 . \tag{5.17}$$

Although finding the best rank-R approximation (R-term CP decomposition) of a tensor is ill-defined in general [101], various numerical algorithms exist in the tensor recovery literature to find a "good" rank-R approximation of a tensor [5, 101]. TeFDiL can employ any of these algorithms to find the R-term CP decomposition, denoted by $\mathrm{CPD}_R(\cdot)$, of $\mathcal{T}^+(\widetilde{\mathbf{y}})$. At the end of each dictionary update stage, the columns of $\mathbf{D} = \sum \bigotimes \mathbf{D}_k^r$ are normalized. Algorithm 2 describes the dictionary update step of TeFDiL.

Algorithm 2 Dictionary Update in TeFDiL for LSR-DL

Require: $\mathbf{Y}$, $\mathbf{X}(t)$, $\mathbf{\Pi}$, r

1: Construct $\mathcal{T}^+(\mathbf{Y}) = \operatorname{vec}^{-1}\left(\mathbf{\Pi}\big(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\big)^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{y}}\right)$
2: $\underline{\mathbf{D}}^\pi \leftarrow \mathrm{CPD}_R(\mathcal{T}^+(\widetilde{\mathbf{y}}))$
3: $\mathbf{D} \leftarrow \operatorname{vec}^{-1}\left(\mathbf{\Pi}^\top \operatorname{vec}(\underline{\mathbf{D}}^\pi)\right)$
4: Normalize columns of $\mathbf{D}$
5: **return** $\mathbf{D}(t+1)$

[3]In our experiments, we add $\delta\mathbf{I}$ to $\mathbf{X}\mathbf{X}^\top$ with a small $\delta > 0$ at every iteration to ensure full-rankness.

5.3.3 OSubDil: An Online LSR-DL Algorithm

Both STARK and TeFDiL are batch methods in that they use the entire dataset for DL in every iteration. This makes them less scalable with the size of datasets due to high memory and per iteration computational cost and also makes them unsuitable for streaming data settings. To overcome these limitations, we now propose an online LSR-DL algorithm termed *Online SubDictionary Learning for structured DL (OSubDil)* that uses only a single data sample (or a small mini-batch) in every iteration (see Algorithm 3). This algorithm has better memory efficiency as it removes the need for storing all data points and has significantly lower per-iteration computational complexity. In *OSubDil*, once a new sample $\underline{\mathbf{Y}}(t+1)$ arrives, its sparse representation $\underline{\mathbf{X}}(t+1)$ is found using the current dictionary estimate $\mathbf{D}(t)$ and then the dictionary is updated using $\underline{\mathbf{Y}}(t+1)$ and $\underline{\mathbf{X}}(t+1)$. The dictionary update stage objective function after receiving the T-th sample is

$$J_T(\{\mathbf{D}_k^r\}) = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{y}(t) - \big(\sum_{r=1}^{R}\bigotimes_{k=1}^{K}\mathbf{D}_k^r\big)\mathbf{x}(t)\|^2. \tag{5.18}$$

We can restate this objective as

$$\begin{aligned} J_T(\{\mathbf{D}_k^r\}) &= \sum_{t=1}^{T} \|\mathbf{Y}_{(k)}(t) - \sum_{r=1}^{R}\mathbf{D}_k^r\mathbf{X}_{(k)}(t)\mathbf{C}_k^r(t)\|_F^2 \\ &= \sum_{t=1}^{T} \|\widehat{\mathbf{Y}}_{(k)}(t) - \mathbf{D}_k^r\mathbf{X}_{(k)}(t)\mathbf{C}_k^r(t)\|_F^2 \\ &= \mathrm{Tr}\left(\mathbf{D}_k^{r\top}\mathbf{D}_k^r\mathbf{A}_k^r(t)\right) - 2\,\mathrm{Tr}\left(\mathbf{D}_k^{r\top}\mathbf{B}_k^r(t)\right) + \text{const.}, \end{aligned}$$

where dropping the iteration index t, $\mathbf{C}_k^r \triangleq \left(\mathbf{D}_K^r \otimes \cdots \otimes \mathbf{D}_{k+1}^r \otimes \mathbf{D}_{k-1}^r \cdots \otimes \mathbf{D}_1^r\right)^\top$, $\widehat{\mathbf{Y}}_{(k)} \triangleq \mathbf{Y}_{(k)} - \sum_{\substack{i=1 \\ i\neq r}}^{R}\mathbf{D}_k^i\mathbf{X}_{(k)}\mathbf{C}_k^i$,

$$\begin{aligned} \mathbf{A}_k^r(t) &\triangleq \sum_{\tau=1}^{t}\mathbf{X}_{(k)}(t)\mathbf{C}_k^r(\tau)\mathbf{C}_k^r(\tau)^\top\mathbf{X}_{(k)}(\tau)^\top, \text{and} \\ \mathbf{B}_k^r(t) &\triangleq \sum_{\tau=1}^{t}\widehat{\mathbf{Y}}_{(k)}(\tau)\mathbf{C}_k^r(\tau)^\top\mathbf{X}_{(k)}(\tau)^\top. \end{aligned} \tag{5.19}$$

To minimize $J_T(\{\mathbf{D}_k^r\})$ with respect to each $\mathbf{D}_k^r$, we take a similar approach as in Mairal et al. [54] and use a (block) coordinate descent algorithm with warm start to update the columns of $\mathbf{D}_k^r$ in a cyclic manner. Algorithm 3 describes the dictionary update step of OSubDil.

Algorithm 3 Dictionary Update in OSubDil for LSR-DL

Require: $\underline{\mathbf{Y}}(t)$, $\{\mathbf{D}_k^r(t)\}$, $\mathbf{A}_k^r(t)$, $\mathbf{B}_k^r(t)$, $\underline{\mathbf{X}}(t)$
for all $r \in [R]$ **do**
  for all $k \in [K]$ **do**
    $\mathbf{C}_k^r \leftarrow \left(\mathbf{D}_K^r \otimes \cdots \otimes \mathbf{D}_{k+1}^r \otimes \mathbf{D}_{k-1}^r \cdots \otimes \mathbf{D}_1^r\right)^\top$
    $\widehat{\mathbf{Y}}_{(k)} \leftarrow \mathbf{Y}_{(k)} - \sum_{\substack{i=1 \\ i \neq r}}^{R} \mathbf{D}_k^i \mathbf{X}_{(k)} \mathbf{C}_k^i$
    $\mathbf{A}_k^r \leftarrow \mathbf{A}_k^r + \mathbf{X}_{(k)} \mathbf{C}_k^r {\mathbf{C}_k^r}^\top {\mathbf{X}_{(k)}}^\top$
    $\mathbf{B}_r^k \leftarrow \mathbf{B}_k^r + \widehat{\mathbf{Y}}_{(k)} {\mathbf{C}_k^r}^\top {\mathbf{X}_{(k)}}^\top$
    for $j = 1, \cdots, p_k$ **do**
      $\mathbf{d}_{k,j}^r \leftarrow \frac{1}{a_{k,jj}^r}(\mathbf{b}_{k,j}^r - \mathbf{D}_k^r \mathbf{a}_{k,j}^r) + \mathbf{d}_{k,j}^r$
    end for
  end for
end for
Normalize columns of $\mathbf{D} = \sum_{r=1}^{R} \bigotimes_{k=1}^{K} \mathbf{D}_k^r$
return $\{\mathbf{D}_k^r(t+1)\}$

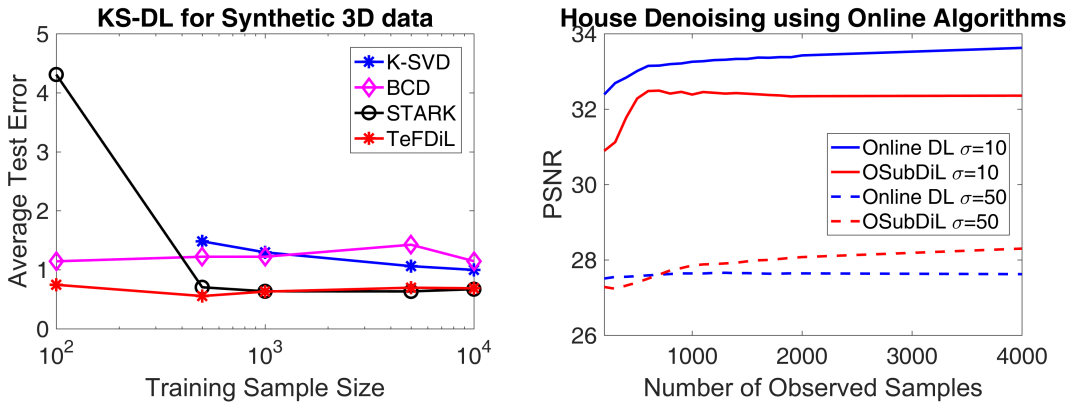
5.4 Numerical Experiments

We evaluate our algorithms on synthetic and real-world datasets to understand the impact of training set size and noise level on the performance of LSR-DL. In particular, we want to understand the effect of exploiting additional structure in representation accuracy and denoising performance. We compare the performance of our proposed algorithms with existing DL algorithms in each scenario and show that in almost every case our proposed LSR-DL algorithms outperform K-SVD. Our results also offer insights into how the size and quality of training data can affect the choice of the proper DL model. Specifically, our experiments on image denoising show that when noise level in data is high, TeFDiL performs best when the separation rank is 1. On the other hand, in low noise regimes, the performance of TeFDiL improves as we increase the separation rank. Furthermore, our synthetic experiments confirm that when the true underlying dictionary follows the KS (LSR) structure, our structured algorithms clearly outperform K-SVD, especially when the number of training samples is very small. This

implies the potential of the LSR-DL model and our algorithms in applications where the true dictionary follows the LSR structure more closely.

Synthetic Experiments: We compare our algorithms to K-SVD[4] (standard DL) as well as a simple block coordinate descent (BCD) algorithm that alternates between updating every coordinate dictionary in problem (5.4). This BCD algorithm can be interpreted as an extension of the KS-DL algorithm [16] for the LSR model. We show how structured DL algorithms outperform the unstructured algorithm K-SVD[4] when the underlying dictionary is structured, especially when the training set is small. We focus on 3rd-order tensor data and we randomly generate a KS dictionary $\mathbf{D} = \mathbf{D}_1 \otimes \mathbf{D}_2 \otimes \mathbf{D}_3$ with dimensions $\mathbf{m} = [2, 5, 3]$ and $\mathbf{p} = [4, 10, 5]$. We select i.i.d samples from the standard Gaussian distribution, $\mathcal{N}(0, 1)$, for the coordinate dictionary elements, and then normalize the columns of the coordinate dictionaries. To generate $\mathbf{x}$, we select the locations of $s = 5$ nonzero elements uniformly at random. The values of those elements are sampled i.i.d. from $\mathcal{N}(0, 1)$. We assume observations are generated according to $\mathbf{Y} = \mathbf{DX}$. In the initialization stage of the algorithms, $\mathbf{D}$ is initialized using random columns of $\mathbf{Y}$ for K-SVD and random columns of the unfoldings of $\mathbf{Y}$ for the structured DL algorithms. Sparse coding is performed using OMP [102]. Due to the invariance of DL to column permutations in the dictionary, we choose reconstruction error as the performance criteria. For $L = 100$, K-SVD cannot be used since $p > L$. Reconstruction errors are plotted in Figure 5.2a. It can be seen that TeFDiL outperforms all the other algorithms.

Real-world Experiments: In this set of experiments, we evaluate the image denoising performance of different DL algorithms on four RGB images, `House`, `Castle`, `Mushroom`, and `Lena`, which have dimensions $256\times256\times3$, $480\times320\times3$, $480\times320\times3$, and $512 \times 512 \times 3$, respectively. We corrupt the images using additive white Gaussian noise with standard deviations $\sigma = \{10, 50\}$. To construct the training data set, we extract overlapping patches of size 8×8 from each image and treat each patch as a 3-dimensional data sample. We learn dictionaries with parameters $\mathbf{m} = [3, 8, 8]$ and $\mathbf{p} = [3, 16, 16]$. In the training stage, we perform sparse coding using FISTA [103] (to reduce training time) with regularization parameter $\lambda = 0.1$ for all algorithms. To perform denoising,



(a) (b)

Figure 5.2: (a) Normalized representation error of various DL algorithms for 3rd-order synthetic tensor data. (b) Performance of online DL algorithms for `House`.

we use OMP with $s = \lceil p/20 \rceil$. To evaluate the denoising performances of the methods, we use the resulting peak signal to noise ratio (PSNR) of the reconstructed images [104]. Table 5.1 demonstrates the image denoising results.

LSR-DL vs Unstructured DL: We observe that STARK outperforms K-SVD in every case when the noise level is high and in most cases when the noise level is low. Moreover, TeFDiL outperforms K-SVD in both low-noise and high-noise regimes for all four images while having considerably fewer parameters (one to three orders of magnitude).

LSR-DL vs KS-DL: We compare our results with KS-DL algorithms SeDiL [14] and BCD [16]. Our LSR-DL methods outperform SeDiL and while BCD has a good performance for $\sigma = 10$, its denoising performance suffers when noise level increases.[4]

Table 5.2 demonstrates the image denoising performance of TeFDiL for `Mushroom` based on the separation rank of TeFDiL. When the noise level is low, performance improves with increasing the separation rank. However, for higher noise level $\sigma = 50$, increasing the number of parameters has an inverse effect on the generalization performance.

Comparison of LSR-DL Algorithms: We compare LSR-DL algorithms BCD,

[4]Note that SeDiL results may be improved by careful parameter tuning.

STARK and TeFDiL. As for the merits of our LSR-DL algorithms over BCD, our experiments show that both TeFDiL and STARK outperform BCD in both noise regimes. In addition, while TeFDiL and STARK can be easily and efficiently used for higher separation rank dictionaries, when the separation rank is higher, BCD with higher rank does not perform well. While STARK has a better performance than TeFDiL for some tasks, it has the disadvantage that it does not output the coordinate dictionaries and does not allow for direct tuning of the separation rank. Ultimately, the choice between these two algorithms will be application dependent. The flexibility in tuning the number of KS terms in the dictionary in TeFDiL (and indirectly in STARK, through parameter λ_1) allows selection of the number of parameters in accordance with the size and quality of the training data. When the training set is small and noisy, smaller separation rank (perhaps 1) results in a better performance. For training sets of larger size and better quality, increasing the separation rank allows for higher capacity to learn more complicated structures, resulting in a better performance.

OSubDil vs Online (Unstructured) DL: Figure 5.2b shows the PSNR for reconstructing `House` using OSubDil and Online DL in [54] based on the number of observed samples. We observe that in the presence of high level of noise, our structured algorithm is able to outperform its unstructured counterpart with considerably less parameters.

5.5 Conclusion

We provided the low-separation-rank dictionary learning model (LSR-DL) to learn structured dictionaries for tensor data. This model bridges the gap between unstructured and separable DL models. We presented two LSR-DL algorithms and showed that they have better generalization performance for image denoising in comparison to unstructured DL algorithm K-SVD [4] and existing KS-DL algorithms SeDiL [14] and BCD [16]. We also presented OSubDil that to the best our knowledge is the first online algorithm that results in LSR or KS dictionaries. We show that OSubDil results in a faster reduction in the reconstruction error in terms of number of observed samples compared to the state-of-the-art online DL algorithm [54] when the noise level in data is high.

Table 5.1: Performance of DL algorithms for image denoising in terms of PSNR

		Unstructured	KS-DL ($r = 1$)			LSR-DL ($r > 1$)		
Image	Noise	K-SVD [4]	SeDiL [14]	BCD [16]	TeFDiL	BCD	STARK	TeFDiL
`House`	$\sigma = 10$	35.6697	23.1895	31.6089	36.2955	32.2952	33.4002	37.1275
	$\sigma = 50$	25.4684	23.6916	24.8303	27.5412	21.6128	27.3945	26.5905
`Castle`	$\sigma = 10$	33.0910	23.6955	32.7592	34.5031	30.3561	37.0428	35.1000
	$\sigma = 50$	22.4184	23.2658	22.3065	24.6670	20.4414	24.4965	23.3372
`Mushroom`	$\sigma = 10$	34.4957	25.8137	33.2797	36.5382	32.2098	36.9443	37.7016
	$\sigma = 50$	22.5495	22.9464	22.8554	22.9284	21.7792	25.1081	22.8374
`Lena`	$\sigma = 10$	33.2690	23.6605	30.9575	34.8854	31.1309	33.8813	35.3009
	$\sigma = 50$	22.5070	23.4207	21.6985	23.4988	19.5989	24.8211	23.1658

Table 5.2: Performance of TeFDiL with various ranks for image denoising in terms of PSNR

Image	Noise	$r = 1$	$r = 4$	$r = 8$	$r = 16$	$r = 32$	K-SVD
`Mushroom`	$\sigma = 10$	36.5382	36.7538	37.4173	37.4906	37.7016	34.4957
	$\sigma = 50$	22.9284	22.8352	22.8384	22.8419	22.8374	22.5495
Number of parameters		265	1060	2120	4240	8480	147456

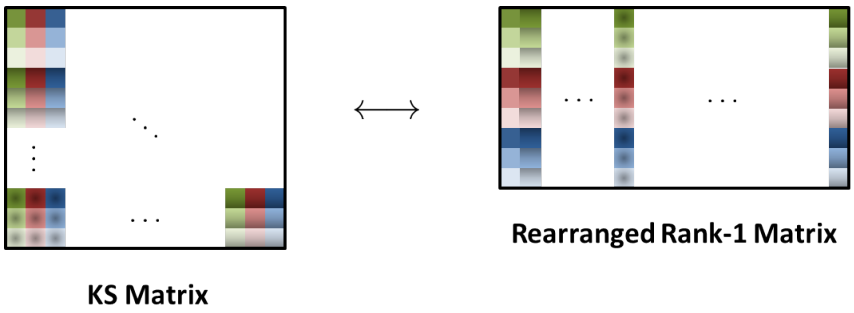


Figure 5.3: Rearranging a KS matrix ($K = 2$) into a rank-1 matrix.

5.6 Appendix

5.6.1 Rearrangement of Kronecker Product to a Low Rank Tensor

To illustrate the procedure that rearranges a KS matrix into a rank-1 tensor, let us first consider $\mathbf{A} = \mathbf{A}_1 \otimes \mathbf{A}_2$. The elements of $\mathbf{A}$ can be rearranged to form $\mathbf{A}^\pi = \mathbf{d}_2 \circ \mathbf{d}_1$, where $\mathbf{d}_k = \text{vec}(\mathbf{A}_k)$ for $k = 1, 2$ [105]. Figure 5.3 depicts this rearrangement for $\mathbf{A}$. Similarly, for $\mathbf{A} = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \mathbf{A}_3$, we can write $\underline{\mathbf{D}}^\pi = \mathbf{d}_3 \circ \mathbf{d}_2 \circ \mathbf{d}_1$, where each frontal slice[5] of the tensor $\underline{\mathbf{D}}^\pi$ is a scaled copy of $\mathbf{d}_3 \circ \mathbf{d}_2$. The rearrangement of $\mathbf{A}$ into $\underline{\mathbf{A}}^\pi$ is performed via a permutation matrix $\mathbf{\Pi}$ such that $\text{vec}(\underline{\mathbf{A}}^\pi) = \mathbf{\Pi}\,\text{vec}(\mathbf{A})$. Given index l of $\text{vec}(\mathbf{A})$ and the corresponding mapped index l' of $\text{vec}(\underline{\mathbf{A}}^\pi)$, our strategy for finding the permutation matrix is to define l' as a function of l. To this end, we first find the corresponding row and column indices (i, j) of matrix $\mathbf{A}$ from the lth element of $\text{vec}(\mathbf{A})$. Then, we find the index of the element of interest on the Kth order rearranged tensor $\underline{\mathbf{A}}^\pi$, and finally, we find its location l' on $\text{vec}(\underline{\mathbf{A}}^\pi)$. Note that the permutation matrix needs to be computed only once in an offline manner, as it is only a function of the dimensions of the factor matrices and not the values of elements of $\mathbf{A}$.

We now describe the rearrangement procedure in detail for the case of KS matrices that are Kronecker product of $K = 3$ factor matrices. Throughout this section, we define a k-th order "tile" to be a scaled copy of $\mathbf{A}_{K-k+1} \otimes \cdots \otimes \mathbf{A}_K$ for $K > 0$. A zeroth order tile is just an element of a matrix.

[5] A slice of a 3-dimensional tensor is a 2-dimensional section defined by fixing all but two of its indices. For example, a frontal slice is defined by fixing the third index.

Kronecker Product of 3 Matrices

In the case of 3rd-order tensors, we take the following steps to find permutation matrix $\mathbf{\Pi}$:

i) Find index (i, j) in $\mathbf{A}$ that corresponds to the l-th element of vec($\mathbf{A}$).

ii) Find the corresponding index (r, c, s) on the third order tensor $\underline{\mathbf{A}}^{\pi}$.

iii) Find the corresponding index l' on vec($\underline{\mathbf{A}}^{\pi}$).

iv) Set $\mathbf{\Pi}(l', l) = 1$.

Let $\mathbf{A} = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \mathbf{A}_3$, with $\mathbf{A} \in \mathbf{R}^{m \times p}$ and $\mathbf{A}_k \in \mathbf{R}^{m_k \times p_k}$ for $k \in [3]$. For the first operation, we have

$$(i, j) = \left(\left\lceil \frac{l}{m} \right\rceil,\ l - \left\lfloor \frac{l-1}{m} \right\rfloor m \right). \tag{5.20}$$

The rearrangement procedure works in the following way. For each element indexed by (i, j) on matrix $\mathbf{A}$, find the 2nd-order tile to which it belongs. Let us index this 2nd-order tile by T_2. Then, find the 1st-order tile (within the 2nd-order tile indexed T_2) on which it lies and index this tile by T_1. Finally, index the location of the element (zeroth-order tile) within this first-order tile by T_0. After rearrangement, the location of this element on the rank-1 tensor is (T_0, T_1, T_2).

In order to find (T_0, T_1, T_2) that corresponds to (i, j), we first find T_2, then T_1, and then T_0. To find T_2, we need to find the index of the 2nd-order tile on which the element indexed by (i, j) lies:

$$T_2 = \underbrace{\left\lfloor \frac{j-1}{p_2 p_3} \right\rfloor}_{S_j^2} m_1 + \underbrace{\left\lfloor \frac{i-1}{m_2 m_3} \right\rfloor}_{S_i^2} + 1, \tag{5.21}$$

where S_j^2 and S_i^2 are the number of the 2nd-order tiles on the left and above the tile to which the element belongs, respectively. Now, we find the position of the element in

this 2nd-order tile:

$$\begin{aligned} i_2 &= i - S_i^2 m_2 m_3 = i - \left\lfloor \frac{i-1}{m_2 m_3} \right\rfloor m_2 m_3, \\ j_2 &= j - S_j^2 p_2 p_3 = j - \left\lfloor \frac{j-1}{p_2 p_3} \right\rfloor p_2 p_3. \end{aligned} \tag{5.22}$$

For the column index, T_1, we have

$$T_1 = \underbrace{\left\lfloor \frac{j_2 - 1}{p_3} \right\rfloor}_{S_j^1} m_2 + \underbrace{\left\lfloor \frac{i_2 - 1}{m_3} \right\rfloor}_{S_i^1} + 1. \tag{5.23}$$

The location of the element on the 1st-order tile is

$$\begin{aligned} i_1 &= i_2 - S_i^1 m_3 = i_2 - \left\lfloor \frac{i_2 - 1}{m_3} \right\rfloor m_3, \\ j_1 &= j_2 - S_j^1 p_3 = j_2 - \left\lfloor \frac{j_2 - 1}{p_3} \right\rfloor p_3. \end{aligned} \tag{5.24}$$

Therefore, T_0 can be expressed as

$$T_0 = (j_1 - 1)\, m_3 + i_1. \tag{5.25}$$

Finally, in the last step we find the corresponding index on $\mathrm{vec}(\underline{\mathbf{A}}^{\pi})$ using the following rule.

$$l' = (T_2 - 1) m_2 m_3 p_2 p_3 + (T_1 - 1) m_3 p_3 + T_0. \tag{5.26}$$

Chapter 6

Computationally Efficient Processing of Tensor Data through Exploitation of Multidimensional Structure

This chapter focuses on the computational advantages of taking the multidimensional structure of tensor data into account for tensor data processing. More specifically, it studies training-based sparse channel estimation in massive MIMO-OFDM systems. In contrast to prior works, the focus here is on the setup in which (training) pilot tones are spread across multiple OFDM symbols. Within this setup, two training models—termed distinct block diagonal (DBD) model and repetitive block diagonal (RBD) model—are investigated. The restricted isometry property, which leads to sparse recovery guarantees, is proven for the DBD model. Further, it is established that the RBD model, through exploitation of its tensor structure, leads to computationally simpler sparse recovery algorithms. Finally, numerical experiments are provided that compare and contrast the channel estimation performance under the two models as a function of the number of pilot tones per OFDM symbol and the total number of OFDM symbols.[1]

6.1 Introduction

Employing multiple antennas in communication systems creates multiple parallel data streams and enhances system reliability [107]. Massive MIMO systems offer many advantages such as increased data throughput and link reliability that are a result of adding extra antennas to MIMO systems [108]. In such systems, coherent signal detection and low bit-error rates rely on the channel state information available at the

[1]The results presented in this chapter have been accepted into the Proceedings of 2019 IEEE International Workshop on Signal Processing Advances in Wireless Communications [106].

receiver. This requires the channel to be periodically estimated at the receiver [17, 107].

The large number of transmit (Tx) and receive (Rx) antennas in massive MIMO systems gives rise to large number of channel parameters, which require considerable spectral resources to estimate them. To reduce spectral resources used for channel estimation, many works exploit the fact that wireless channels associated with a number of scattering environments tend to be highly sparse at high signal space dimension [17, 108]. In this case, training-based channel estimation techniques, which involve transmitting known data to the receiver, can exploit the literature on sparse recovery for reduction in training spectral resources [17, 107, 109].

In this chapter, we study sparse channel estimation of massive MIMO-OFDM channels. Most prior works on sparse channel estimation in MIMO-OFDM systems require the (training) pilot subcarriers (tones) to be interleaved with data subcarriers within one OFDM symbol [17, 107]. But practical systems tend to spread pilot tones across multiple OFDM symbols [110]. While one might anticipate that spreading training resources across frequency and time will result in the same channel estimation performance as using the same number of resources in one OFDM symbol, no prior work has formally investigated this problem to the best of our knowledge. Specifically, let N_t denote the number of OFDM symbols and let N_f be the number of OFDM pilot tones per OFDM symbol. Then the total number of training resources is $N_{tr} = N_t N_f$ and the question we want to address is: *Does the performance of sparse channel estimation in massive MIMO-OFDM systems depend on N_{tr} alone or is it also a function of N_t and N_f?*

In order to address this question, we focus on two models for training in massive MIMO-OFDM systems. In the first model, termed the *distinct block diagonal* (DBD) model, we assume independent training data are transmitted over different pilot tones. Under this model, we show that a channel with no more than S non-zero parameters can be reliably recovered from training observations as long as $N_{tr} = \Omega(S \log^2 S \log^3 p)$, where p denotes the number of channel parameters per Rx antenna. While this result suggests that estimation of (massive) MIMO-OFDM channels is largely a function of the total number of training spectral resources N_{tr}, we rush to add that this is just

a sufficient condition and it comes with a few caveats that are discussed later in the paper.

The second training model discussed in this chapter is termed the *repetitive block diagonal* (RBD) model, in which same training data are transmitted across different pilot tones. The motivation for this model comes from the need to reduce computational and storage complexity at the receiver in downlink settings. Consider, for instance, the setup involving 64 Tx antennas, 4 Rx antennas, and 320 delay taps per Tx-Rx pair. This results in an 81,920-dimensional channel estimation problem at the receiver, requiring large computational and storage resources. The RBD model, however, can be formulated as a Tucker decomposition of the observations [111]. In this case, we show the channel coefficient "tensor" can be recovered using the sparse tensor recovery technique referred to as Kronecker-OMP [28], which has similar performance as the classical orthogonal matching pursuit (OMP) algorithm [112], but has significantly less computational complexity and memory requirements. While we do not derive theoretical guarantees for the RBD model, we provide numerical experiments to compare its performance to that of the DBD model.

In our numerical experiments, we also study the impact of different values of N_t, N_f, and N_{tr} on the performance of both DBD and RBD models. We further investigate the use of overcomplete DFT bases, instead of the canonical bases, to model the angles of arrival (AoA) and angles of departure (AoD) in MIMO channels. Our results show that this leads to enhanced channel estimation. This suggests that data-driven bases can be learned using methods such as dictionary learning to achieve improved channel estimation performance [113].

The rest of this chapter is organized as follows. In Section 6.2, we formulate the MIMO-OFDM channel estimation problem and define our two models. In Section 6.3, we provide recovery guarantees for sparse channels recovered from observations following the DBD model. In Section 6.4, we provide numerical experiments to demonstrate the performance of both modeling techniques. Finally, we conclude the chapter in Section 6.5.

6.2 Problem Formulation

Consider a massive MIMO-OFDM system communicating over a broadband multipath channel $\mathcal{G}$. Let N_T and N_R denote the number of Tx and Rx antennas, respectively, that are half-wavelength spaced linear arrays. Moreover, given channel bandwidth W and symbol duration T, denote $N_0 = WT$ as the temporal signal space dimension, i.e, the number of OFDM subcarriers. Assuming $W \gg 1/\tau_{\max}$, where $\tau_{\max}$ denotes the maximum delay spread of the channel, the frequency response of channel $\mathcal{G}$ can be expressed as[2]

$$\mathcal{G}(f) = \sum_{n=1}^{N_p} \beta_n \mathbf{a}_R(\theta_{R,n}) \mathbf{a}_T^{\mathsf{H}}(\theta_{T,n}) e^{-j2\pi\tau_n f}, \tag{6.1}$$

where N_p is the number of physical paths and $\mathbf{a}_R(\theta_{R,n})$ and $\mathbf{a}_T(\theta_{T,n})$ are the receive and transmit steering vectors, respectively. Here, β_n, $\theta_{R,n}$, $\theta_{T,n}$, and τ_n denote the complex path gain, AoA, AoD, and delay associated with the n-th path, respectively. The physical channel model (6.1) involves a large number of parameters. This motivates a virtual channel representation $\underline{\mathbf{G}}$ of $\mathcal{G}$ that can compactly and linearly model interactions between the Tx and Rx antennas. This involves a discretized approximation of $\mathcal{G}$ by sampling the angle-delay space at Nyquist rate to obtain a 3rd-order tensor $\underline{\mathbf{G}} \in \mathbb{R}^{N_R \times N_T \times N_0}$ that can be expressed via the Tucker decomposition [19] as

$$\underline{\mathbf{G}} = \underline{\mathbf{H}} \times_1 \mathbf{A}_R \times_2 \mathbf{A}_T \times_3 \mathbf{A}_F, \tag{6.2}$$

where $\underline{\mathbf{H}} \in \mathbb{C}^{N_R \times N_T \times L}$ denotes the virtual channel coefficient tensor with $L \triangleq \lceil W\tau_{\max} \rceil + 1$, $\mathbf{A}_R \in \mathbb{C}^{N_R \times N_R}$, $\mathbf{A}_T \in \mathbb{C}^{N_T \times N_T}$, and $\mathbf{A}_F \in \mathbb{C}^{N_0 \times L}$ are the canonical DFT bases associated with AoA, AoD, and delay spread that are used to map $\underline{\mathbf{H}}$ to $\underline{\mathbf{G}}$ [111]. In particular, each element of $\underline{\mathbf{H}}$ can be expressed in terms of the physical propagation

[2]We do not consider the Doppler spread in the scope of this work for simplicity.

path parameters as

$$\underline{\mathbf{H}}(i,k,l) = \sum_{n=1}^{N_p} \beta_n f_{N_R}(\frac{i}{N_R} - \theta_{R,n}) f^*_{N_T}(\frac{k}{N_T} - \theta_{T,n}) \operatorname{sinc}(l - W\tau_n), \tag{6.3}$$

where $f_{N_R}(\theta_R)$ and $f_{N_T}(\theta_T)$ denote the Tx and Rx smoothing kernels defined as $f_N(\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N-1} e^{-j2\pi i\theta}$, and $\operatorname{sinc}(x) \triangleq \sin(\pi x)/\pi x$.

In wideband scenarios, majority of the entries of $\underline{\mathbf{H}}$ tend to be below the noise floor. Our goal is to estimate the resulting "sparse" (or approximately sparse) channel coefficient tensor $\underline{\mathbf{H}}$ using pilot training sequences transmitted over pilot subcarriers spread across N_t OFDM symbols (we assume that the channel stays constant over N_t OFDM symbols). We specifically focus on the setting in which the same set of N_f (out of N_0) subcarriers per OFDM symbol are reserved for training purposes, resulting in a total of $N_{tr} = N_t N_f$ pilot tones. Let $\mathcal{N}_f$ denote the indices of the pilot tones per OFDM symbol and $\mathbf{F} \in \mathbb{R}^{N_f \times N_0}$ denote the subcarrier selection matrix that is comprised of rows of $\mathbf{I}_{N_0}$ corresponding to $\mathcal{N}_f$. Then, each slice of training data $\underline{\mathbf{Y}} \in \mathbb{C}^{N_R \times N_t \times N_f}$ observed at the Rx antennas after N_t symbols can be expressed as

$$\underline{\mathbf{Y}}(:,:,i) = \underline{\mathbf{H}} \times_1 \mathbf{A}_R \times_2 \mathbf{X}_i \mathbf{A}_T \times_3 \mathbf{f}_i \mathbf{A}_F + \underline{\mathbf{W}}(:,:,i), \tag{6.4}$$

where $i \in [N_f]$, $\mathbf{f}_i$ denotes the i-th row of $\mathbf{F}$, and $\underline{\mathbf{W}} \in \mathbb{C}^{N_R \times N_t \times N_f}$ is the additive noise tensor. Here, $\mathbf{X}_i = \mathbf{X}_i^0 \mathbf{B}$ denotes the pilot sequence transmitted over the i-th subcarrier in which $\mathbf{X}_i^0 = \{\pm 1\}^{N_t \times N_t}$ is a square orthogonal matrix and $\mathbf{B} \in \mathbb{C}^{N_t \times N_T}$ is a beamforming matrix that has unit-modulus entries with random phases. This ensures the matrix $\mathbf{X}_i \mathbf{A}_T$ will have similar norm columns.

We refer to the training model described by (6.4) as the distinct block diagonal (DBD) model. This model can be simplified further by assuming that $\mathbf{X}_i \triangleq \mathbf{X}$ for all $i \in [N_f]$, which reduces (6.4) to

$$\underline{\mathbf{Y}} = \underline{\mathbf{H}} \times_1 \mathbf{A}_R \times_2 \mathbf{X}\mathbf{A}_T \times_3 \mathbf{F}\mathbf{A}_F + \underline{\mathbf{W}}. \tag{6.5}$$

Using properties of the Tucker decomposition [19], we can rewrite (6.5) as

$$\text{vec}(\underline{\mathbf{Y}}) = (\mathbf{F}\mathbf{A}_F \otimes \mathbf{X}\mathbf{A}_T \otimes \mathbf{A}_R)\,\text{vec}(\underline{\mathbf{H}}) + \text{vec}(\underline{\mathbf{W}}), \tag{6.6}$$

where $\otimes$ denotes the matrix Kronecker product and $\text{vec}(\underline{\mathbf{Y}})$ denotes the vectorized version of $\underline{\mathbf{Y}}$. We refer to this training model as the repetitive block diagonal (RBD) model.

Our focus in this chapter is addressing the questions: *Can we guarantee recovery of sparse* $\underline{\mathbf{H}}$ *under the DBD and RBD models? What are the computational advantages of using the RBD model compared to the DBD model?* We first address the first question theoretically for the DBD model in the next section. Afterwards, we focus on the numerical aspects of this question for both DBD and RBD models in Section 6.4. In particular, due to the fact that (6.5) follows the Tucker decomposition, the RBD model allows recovery of $\underline{\mathbf{H}}$ using tensor recovery techniques. Specifically, Kronecker-OMP is a method introduced in [28] that does not require explicit computation of the Kronecker-structured measurement matrix $(\mathbf{F}\mathbf{A}_F \otimes \mathbf{X}\mathbf{A}_T \otimes \mathbf{A}_R)$ in (6.6), thus facilitating recovering of $\underline{\mathbf{H}}$ using less computation complexity and memory requirements compared to regular OMP.

In the next section, we show that under the DBD model in (6.4), $\underline{\mathbf{H}}$ is recoverable under certain conditions on the measurement matrix.

6.3 Sparse Channel Estimation Under the DBD Model

Let us consider the linear observation model $\mathbf{y} = \mathbf{A}\mathbf{h} + \text{vec}(\underline{\mathbf{W}})$ in which $\mathbf{h}$ is S-sparse (i.e., has no more than S non-zero entries). We first describe a property that is essential for recovering $\mathbf{h}$ from $\mathbf{y}$.

Proposition 6.1. *Let* $\mathbf{A} \in \mathbb{C}^{nk\times p}$ *be a matrix with unit-norm columns. To ensure reliable recovery of an* S*-sparse* $\mathbf{h} \in \mathbb{C}^p$ *from* $\mathbf{y}$*,* $\mathbf{A}$ *has to satisfy the restricted isometry property (*RIP*) of order* S*, i.e.,* $\mathbf{A} \in \mathsf{RIP}(S, \delta_S)$ *with* $\delta_S \in (0,1)$ *if for all* S*-sparse* $\mathbf{h}$*,*

$$(1-\delta_S)\|\mathbf{h}\|_2^2 \le \|\mathbf{A}\mathbf{h}\|_2^2 \le (1+\delta_S)\|\mathbf{h}\|_2^2. \tag{6.7}$$

Notice that $\mathbf{A} \in \mathsf{RIP}(S, \delta_S)$ if we have

$$\max_{\mathcal{T} \subset [p], |\mathcal{T}| \leq S} \left\| \mathbf{A}_{\mathcal{T}}^{\mathsf{H}} \mathbf{A}_{\mathcal{T}} - \mathbf{I}_{|\mathcal{T}|} \right\|_2 \leq \delta_S, \tag{6.8}$$

where $\mathbf{A}_{\mathcal{T}}$ denotes the matrix consisting of columns of $\mathbf{A}$ with indices $\mathcal{T}$ and $\mathbf{I}_{|\mathcal{T}|}$ denotes the identity matrix of size $|\mathcal{T}| \times |\mathcal{T}|$. Using the non-negative function $\|.\|_{\mathcal{T},S} : \mathbb{C}^{p \times p} \to [0, \infty)$ that is defined as $\|\mathbf{P}\|_{\mathcal{T},S} \triangleq \max_{\mathcal{T} \subset [p], |\mathcal{T}| \leq S} \|\mathbf{P}_{\mathcal{T} \times \mathcal{T}}\|_2$, where $\mathbf{P}_{\mathcal{T} \times \mathcal{T}}$ is a submatrix of $\mathbf{P}$ constructed by collecting entries of $\mathbf{P}$ with indices in the set $\mathcal{T} \times \mathcal{T}$, (6.8) can be restated as

$$\left\| \mathbf{A}^{\mathsf{H}} \mathbf{A} - \mathbf{I}_p \right\|_{\mathcal{T},S} \leq \delta_S. \tag{6.9}$$

We now provide a theorem that shows that a special class of *structured* matrices satisfies the RIP under certain conditions. The ensuing discussion then relates this class of matrices to the observations arising within the DBD model.

Theorem 6.1. *Let $\mathbf{U} \in \mathbb{C}^{p \times p}$ be a unitary matrix. Define $\mathcal{X} \triangleq \left\{ x_{i,i'} \right\}$, where $i \in [mk]$, $i' \in [k']$, and $k' \triangleq p/m$ is an integer factor of p, to be a generating sequence whose elements are independent realizations of Rademacher random variables taking values ± 1 with probability $1/2$. Let $\mathbf{R} \in \mathbb{R}^{mk \times p}$ be a block diagonal row-mixing matrix with $mk \leq p$, defined as*

$$\mathbf{R} \triangleq \begin{bmatrix} \mathbf{R}_1 & 0 & \dots & 0 \\ 0 & \mathbf{R}_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{R}_m \end{bmatrix}, \tag{6.10}$$

where

$$\mathbf{R}_i \triangleq \begin{bmatrix} x_{(i-1)k+1,1} & \dots & x_{(i-1)k+1,k'} \\ \vdots & \ddots & \vdots \\ x_{ik,1} & \dots & x_{ik,k'} \end{bmatrix}. \tag{6.11}$$

Next, define $\mathbf{\Phi} \triangleq \mathbf{RU}$. *Further, given a subset* Ω *of cardinality* $|\Omega| = n$ *chosen uniformly at random without replacement from* $[m]$, *define* Ω' *of cardinality* $|\Omega'| = nk$ *with elements* $\Omega' = \{(i-1)k + j, i \in \Omega, j \in [k]\}$. *Also, let* $\mathbf{A} \in \mathbb{C}^{nk \times p}$ *be the result of sampling* nk *rows of* $\mathbf{\Phi}$ *with indices in* Ω' *and normalizing the resulting columns by* $\sqrt{m/(kn)}$. *Finally, define* $\mu_{\mathbf{U}} \triangleq \sqrt{p}\max_{i,j}|u_{ij}|$ *as the coherence of* $\mathbf{U}$. *Then, for each integer* p, $S > 2$, *and for any* $z > 1$ *and any* $\delta_S \in (0,1)$, *there exist positive constants* c_1 *and* c_2 *such that if* $nk \geq c_1 z \mu_{\mathbf{U}}^2 S \log^2 S \log^3 p$, *then* $\mathbf{A}$ *satisfies* $\mathsf{RIP}(S, \delta_S)$ *with probability higher than* $1 - 20\max\left\{\exp\left(-c_2\delta_S^2 z\right), p^{-1}\right\}$.

Similar to the proof provided by Bajwa et al. [114], we prove this theorem by first assuming that the block sampling variables in Ω follow Bernoulli distribution, and then translate the results for uniform distribution. To this end, let $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^m$ be independent Bernoulli random variables taking value 1 with probability n/m and let $\Omega \triangleq \{i : \xi_i = 1\}$. Also, define $\boldsymbol{\eta} = \{\eta_j\}_{j=1}^{mk} = \boldsymbol{\xi} \otimes \mathbf{1}_k$ and $\Omega' = \{j : \eta_j = 1\}$. We then have the following lemmas. In all lemmas, it is assumed that $\mathbf{A}$ is a structurally-subsampled unitary matrix, as defined in Theorem 6.1, generated from $\mathbf{\Phi}$ according to the Bernoulli sampling model.

Lemma 6.1. *We have* $\mathbb{E}\left[\mathbf{A}^{\mathsf{H}}\mathbf{A}\right] = \mathbf{I}_p$.

Proof. The proof follows from steps similar to those in [115, Lemma 3.10] after some algebraic manipulations. □

Lemma 6.2. *For any integer* $p > 2$ *and any* $r \in [2, 2\log p]$, *we have*

$$\left(\mathbb{E}\left[\|\mathbf{A}\|_{\max}^r\right]\right)^{1/r} \leq \sqrt{\frac{m}{nk}}\left(\mathbb{E}\left[\|\mathbf{\Phi}\|_{\max}^r\right]\right)^{1/r} \leq \sqrt{\frac{16\mu_{\mathbf{U}}^2 \log p}{nk}}. \tag{6.12}$$

Proof. The proof relies on the Khintchine inequality [116, Lemma 4.1], and follows similar steps as in [115, Lemma 3.13]. □

Lemma 6.3. *For any integer* $p > 2$ *and any* $\varepsilon \in (0,1)$, *we have* $\mathbb{E}\left[\left\|\mathbf{A}^{\mathsf{H}}\mathbf{A} - \mathbf{I}_p\right\|_{\mathcal{T},S}\right] \leq$

$k\varepsilon$ provided

$$nk \geq c_3 \varepsilon^{-2} \mu_{\mathbf{U}}^2 S \log^2 S \log^3 p, \tag{6.13}$$

for some positive constant c_3.

Proof. We have

$$\mathbb{E}\left[\left\|\mathbf{A}^{\mathsf{H}}\mathbf{A} - \mathbf{I}_p\right\|_{\mathcal{T},S}\right] \overset{(a)}{\leq} \sum_{l=1}^{k} \mathbb{E}\left[\left\|\mathbf{A}_l^{\mathsf{H}}\mathbf{A}_l - \frac{1}{k}\mathbf{I}_p\right\|_{\mathcal{T},S}\right], \tag{6.14}$$

where $\mathbf{A}_l$ denotes the matrix comprised of rows of $\mathbf{A}$ with indices $\{(i-1)k+l\}_{i=1}^{m}$ and (a) follows from Jensen's inequality since $\|.\|_{\mathcal{T},S}$ is a norm [115]. We can show that $\mathbb{E}\left[\left\|\mathbf{A}_l^{\mathsf{H}}\mathbf{A}_l - \frac{1}{k}\mathbf{I}_p\right\|_{\mathcal{T},S}\right] \leq \varepsilon$ using similar steps as in[115, Lemma 3.14] that takes advantage of the Rudelson-Vershynin inequality [117, Lemma 3.8]. □

Proof of Theorem 6.1. A result from [118, Section 2.3] states that if subsampled matrices from a certain class satisfy RIP with probability exceeding $1-\zeta$ for the Bernoulli sampling model, then they also satisfy RIP with probability exceeding $1-2\zeta$ for the uniformly-at-random sampling model. It can be shown that this result holds for the case of our block Bernoulli and uniformly-at-random sampling models as well. Hence, it is sufficient to show that $\mathbf{A}$ satisfies $\mathsf{RIP}(\delta_S, S)$ for the block Bernoulli sampling model. We next define

$$\mathbf{Y}_i \triangleq \frac{m}{nk}\xi_i \mathbf{\Phi}_i^{\mathsf{H}}\mathbf{\Phi}_i - \frac{1}{m}\mathbf{I}_p, \qquad \widetilde{\mathbf{Y}}_i \triangleq \frac{m}{nk}\left(\xi_i \mathbf{\Phi}_i^{\mathsf{H}}\mathbf{\Phi}_i - \xi_i' \mathbf{\Phi}_i'^{\mathsf{H}}\mathbf{\Phi}_i'\right), \tag{6.15}$$

for $i \in [m]$. Here, $\mathbf{\Phi}_i$ denotes the matrix comprised of rows of $\mathbf{\Phi}$ with indices $\{(i-1)k+l\}_{l=1}^{k}$, ξ_i' and $\mathbf{\Phi}_i'$ are independent copies of ξ_i and $\mathbf{\Phi}_i$, and hence, $\sum_{i=1}^{m}\widetilde{\mathbf{Y}}_i$ is a symmetric version of $\sum_{i=1}^{m}\mathbf{Y}_i$. Defining $\widetilde{\mathbf{Y}} \triangleq \left\|\sum_{i=1}^{m}\widetilde{\mathbf{Y}}_i\right\|_{\mathcal{T},S}$ and $\mathbf{Y} \triangleq \|\sum_{i=1}^{m}\mathbf{Y}_i\|_{\mathcal{T},S}$, from [116], we have for all $u > 0$:

$$\mathbb{E}\big[\widetilde{\mathbf{Y}}\big] \leq 2\mathbb{E}\left[\mathbf{Y}\right], \qquad \mathbb{P}\left[\mathbf{Y} > 2\mathbb{E}\left[\mathbf{Y}\right] + u\right] \leq 2\mathbb{P}\big[\widetilde{\mathbf{Y}} > u\big]. \tag{6.16}$$

Hence, from Lemma 6.3, $\mathbb{E}\big[\widetilde{\mathbf{Y}}\big] \leq 2k\varepsilon$. We can use Lemma 6.2 and Markov's inequality to show that with probability exceeding $1-2p^{-1}$, $\max_i \left\|\widetilde{\mathbf{Y}}_i\right\|_{\mathcal{T},S} \leq 2SB_1$, where $B_1 \triangleq \frac{16e\mu_{\mathbf{U}}^2 \log p}{n}$. Conditioned on the event $F \triangleq \left\{\max_i \left\|\widetilde{\mathbf{Y}}_i\right\|_{\mathcal{T},S} \leq 2SB_1\right\}$, using Lemma 6.3 and the Ledoux-Talagrand inequality [117, Lemma 3.10], if (6.13) is satisfied, then for any integer $r \geq q$, any $t > 0$, some absolute constant $c_4 > 0$, and any $\varepsilon \in (0, 1/k)$:

$$\mathbb{P}\big[\widetilde{\mathbf{Y}} \geq 16qk\varepsilon + 4rSB_1 + t|F\big] < \frac{c_4^r}{q^r} + 2\exp\left(\frac{-t^2}{1024qk^2\varepsilon^2}\right). \tag{6.17}$$

Next, choose $q = \lceil ec_4 \rceil$, $t = 32\sqrt{q}\zeta k\varepsilon$ and $r = \lceil \frac{t}{2SB_1} \rceil$ for some $\zeta > 1$, and define $c_1 \triangleq \max\left\{e\sqrt{q}, c_3\right\}$. Given $\mathbb{P}(F^c) \leq 2p^{-1}$, if $nk \geq c_1\varepsilon^{-2}\mu_{\mathbf{U}}^2 S\log^2 S\log^3 p$, then $r \geq q$ and

$$\mathbb{P}\big[\widetilde{\mathbf{Y}} \geq (16q + 96\sqrt{q}\zeta)k\varepsilon\big] < \exp\left(-\frac{\sqrt{q}\zeta\varepsilon kn}{3\mu_{\mathbf{U}}^2 S\log p}\right) + 2\exp(-\zeta^2) + 2p^{-1}. \tag{6.18}$$

We can translate this result for $\mathbf{Y}$ using (6.16). If (6.13) is satisfied, then $\mathbb{E}\left[\mathbf{Y}\right] \leq k\varepsilon$ from Lemma 6.3. In this case, we get

$$\begin{aligned}\mathbb{P}\left[\mathbf{Y} \geq (2 + 16q + 96\sqrt{q}\zeta)k\varepsilon\right] &< 2\exp\left(-\frac{\sqrt{q}\zeta\varepsilon kn}{3\mu_{\mathbf{U}}^2 S\log p}\right) + 4\exp(-\zeta^2) + 4p^{-1} \\ &\overset{(a)}{<} 10\max\left\{\exp\left(-c_2\delta_S^2 z\right), p^{-1}\right\},\end{aligned} \tag{6.19}$$

where (a) follows from defining $c_5 \triangleq 2 + 16q + 96\sqrt{q}$ (which implies $c_5\zeta k\varepsilon > (2 + 16q + 96\sqrt{q}\zeta)k\varepsilon$), choosing $\zeta = \frac{\delta_S}{c_5 k\varepsilon}$, and denoting $c_2 \triangleq 1/c_5$ and $z \triangleq 1/(k\varepsilon)^2$. □

6.3.1 Discussion

Theorem 6.1 implies that if $nk = \Omega(\mu_{\mathbf{U}} S\log^2 S\log^3 p)$, $\mathbf{A}$ will satisfy $\mathsf{RIP}(S, \delta_S)$ with $\delta_S = \Omega(k\varepsilon)$ for an appropriately small $\varepsilon \in (0,1)$ and an S-sparse $\mathbf{h}$ is recoverable from $\mathbf{y}$ with high probability. Notice however that for large values of k, $\delta_S > 1$ and Theorem 6.1 will not hold. This restriction is a limitation of our proof technique.

Connecting Theorem 6.1 to the MIMO-OFDM observation model in (6.4), let $\mathbf{y}_r \in \mathbb{C}^{N_t N_f}$, $r \in [N_R]$, denote the vectorized observation received at antenna r. Also, let the

indices of the N_f pilot tones be selected uniformly at random from $[N_0]$. In this case, $\mathbf{y}_r$ can be divided into N_f blocks: $\mathbf{y}_r = \begin{bmatrix}\mathbf{y}_r(1)^\top & \dots & \mathbf{y}_r(N_f)^\top\end{bmatrix}^\top$. We can write

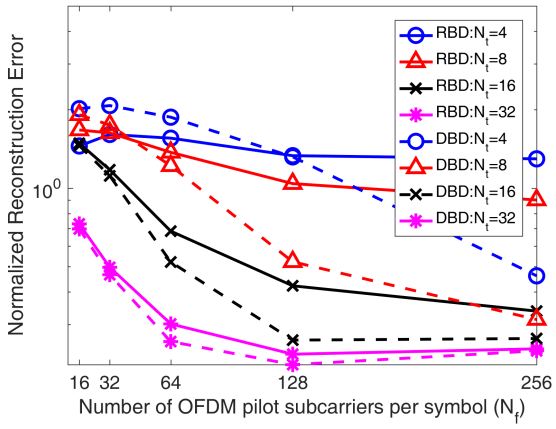
$$\mathbf{y}_i(r) = \mathbf{X}_i(\mathbf{f}_i\mathbf{A}_F \otimes \mathbf{A}_T)\mathbf{h}_r + \mathbf{w}_i(r),\ i \in [N_f], \tag{6.20}$$

where $\mathbf{h}_r \in \mathbb{C}^{N_T L}$ is a vectorized version of channel coefficients $\underline{\mathbf{H}}(r,k,l)$, where $k \in [N_T]$ and $l \in [L]$. This corresponds to the observation model in Theorem 6.1 where $\mathbf{R}_i = \mathbf{X}_i$, $\mathbf{U}$ consists of stacking $\frac{1}{\sqrt{N_0}}(\mathbf{a}_{F,i} \otimes \mathbf{A}_T)$ on top of each other, where $\mathbf{a}_{F,i}$ denotes the ith row of $\mathbf{A}_F$ for $i \in [N_0]$, $k = N_t$, $n = N_f$, and $p = N_T L$. Here, $\mu_{\mathbf{U}} = 1$. This means that for reliable recovery of $\mathbf{h}_r$, $N_t N_f = \Omega(S \log^2 S \log^3 N_T L)$ has to be satisfied. In comparison to the result provided by Bajwa et al. [114] for the case of $N_t = 1$ that requires scaling of $N_f = \Omega(S \log^2 S \log^3 N_T L)$, it can be seen that the total number of parameters in $\mathbf{X}_i$, i.e. $N_{tr} = N_t N_f$, is the determining factor for reliable recovery of $\mathbf{h}_r$ in our setup. However, note that the theorem does not hold for large values of N_t since in that case $\delta_S > 1$.

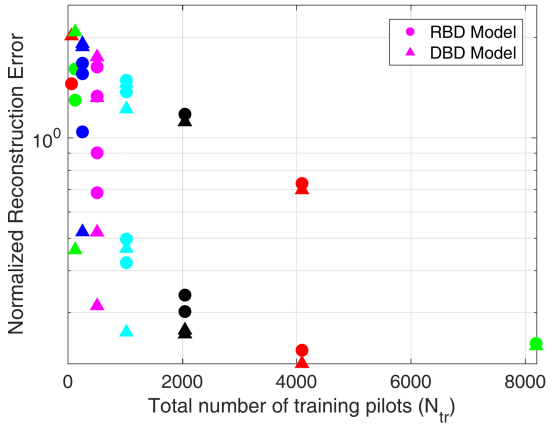
Our discussion so far has been focused on sufficient conditions. In the next section, we show numerically that the sparse channel estimation performance actually depends on individual values of N_t and N_f as well as on N_{tr}.

6.4 Numerical Results

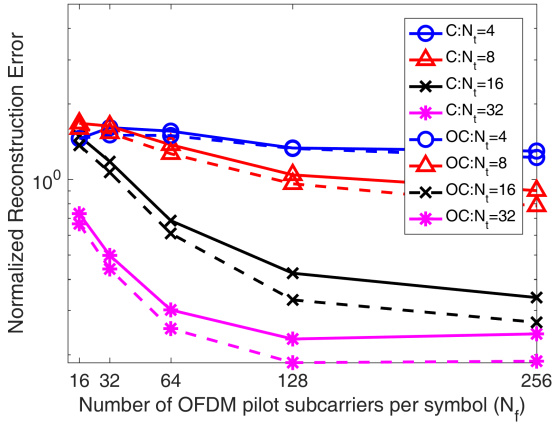
In this section, we evaluate the performance of sparse channel estimation under the DBD and RBD models in terms of the number of OFDM pilot subcarriers per symbol N_f and the number of symbols N_t. The experimental setup corresponds to $N_R = 4$, $N_T = 64$, $N_0 = 1024$, $N_p = 200$. We assume $W = 25.12$MHz and τ_n's are uniformly distributed over $[0, 12.7\mu\text{sec}]$, resulting in $L = 320$. Moreover, $(\theta_{R,n}, \theta_{T,n})$'s are uniformly distributed over $[-1/2, 1/2] \times [-1/2, 1/2]$ and β_n's follow the normal distribution. We use Gaussian noise with standard deviation $\sigma = 0.2\sqrt{2}$. We select the set of pilot subcarriers $\mathcal{N}_f$ uniformly-at-random from $[N_0]$. We generate random $\mathbf{X}_i$'s according to the description in Section 6.2. We generate channel realization coefficients according to (6.3) and conduct experiments for $N_t = [4, 8, 16, 32]$ and $N_f = [16, 32, 64, 128, 256]$.



(a)



(b)



(c)

Figure 6.1: Normalized reconstruction error for DBD and RBD models as a function of (a) N_f and (b) N_{tr}. In (c), we plot the normalized reconstruction error for complete (C) and overcomplete (OC) AoA and AoD bases (RBD model only).

We use OMP and Kronecker-OMP with sparsity level $S = 1000$ to reconstruct $\underline{\mathbf{H}}$ from noisy observations $\underline{\mathbf{Y}}$ for DBD and RBD models, respectively.

We evaluate the channel estimation performance via the normalized reconstruction error, i.e., $\frac{\|\underline{\mathbf{G}}-\widehat{\underline{\mathbf{G}}}\|_F^2}{\|\underline{\mathbf{G}}\|_F^2}$. In all experiments, we average the error over 100 Monte Carlo experiments for random channel, additive noise, and training pilot realizations.

We conduct two sets of experiments. In both sets, the reconstruction error is plotted against N_f for various N_t's. In the first set, we compare the performance of channel estimation using the DBD and the RBD models. Figure 6.1a shows the reconstruction performance for both models (solid lines represent RBD model while dotted lines represent DBD model). For both models, it can been seen that lower error levels are achieved by increasing N_f for all N_t's. We also achieve better reconstruction when we choose a larger N_t. It can also be seen that although the DBD model outperforms the RBD model for smaller values of N_t and N_f, their performance is similar for larger values, especially for $N_f = 256$. This shows that given sufficient training pilot tones, both models have a similar performance and one can use the RBD model to take advantage of Kronecker-OMP to reduce storage costs and required computational resources at the Rx.

Figure 6.1b shows the error for both training models as a function of the total number of pilot tones, $N_{tr} = N_t N_f$. While it is clear that the general trend is downward based on N_{tr}, it is observed that N_{tr} is not the only determining factor and values of N_f and N_t individually matter as well in determining the error.

In the second set of experiments, we compare the performance of channel estimation using complete (C) and overcomplete (OC) bases under the RBD model. We use factor matrices $\mathbf{A}_R$, $\mathbf{X}\mathbf{A}_T$, and $\mathbf{F}\mathbf{A}_F$ to form the measurement matrix in the complete case (solid lines in Figure 6.1c) and we use overcomplete DFT matrices instead of $\mathbf{A}_R$ and $\mathbf{A}_T$ in the overcomplete setup (dotted lines in figure 6.1c). It can be observed in Figure 6.1c that the use of overcomplete DFT bases results in a reduction in the reconstruction error. This suggests that perhaps these matrices can be carefully designed using dictionary learning techniques similar to those in [88, 113] for enhanced reconstruction performance.

6.5 Conclusion

In this chapter, we focused on the computational aspects of multidimensional processing of tensor data and we studied the sparse channel estimation problem for (massive) MIMO-OFDM systems. Here, the underlying channel can be modeled as a tensor with three modes (angle of arrival, angle of departure, and delay spread). We introduced the distinct block diagonal model for training data and obtained theoretical guarantees for channel recovery based on number of training pilot tones. Moreover, we studied the repetitive block diagonal model for training data that results in a Tucker decomposition for the observations. This formulation allows recovery of channel coefficients using sparse tensor recovery techniques that use less computational measures and memory compared to traditional recovery techniques.

We further provided a comparison of the performance of the two models via numerical experiments. While our theory states that the total number of parameters determine the channel estimation performance, our numerical experiments show that the performance is also a function of the number of OFDM symbols and pilot tones. Consequently, there is a lot more that needs to be understood about the performance of these models. Future work includes providing formal guarantees for the repetitive block diagonal model using proof techniques similar to those in [119]. Our perspectives also include the use of dictionary learning techniques to improve the channel estimation performance for more challenging scenarios.

Chapter 7

Conclusion and Future Work

In this dissertation, we provided key results for the fundamental limits of DL methods that explicitly account for the multidimensional structure of tensor data through KS dictionaries. We also proposed structured DL algorithms for efficient tensor data representation. Finally, we investigated the computational advantages of using tensor recovery techniques over vectorized methods for channel estimation in MIMO-OFDM systems. In this chapter, we summarize our results and discuss related open problems and future work.

7.1 Kronecker Structured Dictionary Learning for Tensor Data

In Chapter 3, we followed an information-theoretic approach to provide lower bounds for the worst-case MSE of KS dictionaries that generate Kth-order tensor data. We established that estimating a KS dictionary comprising of K coordinate dictionaries with dimensions $\{m_k \times p_k\}$ requires a number of samples that needs to grow only linearly with the sum of the sizes of the component dictionaries, i.e., $\sum_{k\in[K]} m_k p_k$. We also demonstrated that for a special case of $K = 2$, there exists an estimator whose MSE meets one of our derived lower bounds. While our analysis is local in the sense that we assume the true dictionary belongs in a local neighborhood with known radius around a fixed reference dictionary, the derived minimax risk effectively becomes independent of this radius for sufficiently large neighborhood radius.

Furthermore, in Chapter 4, we derived sufficient conditions for local recovery of coordinate dictionaries comprising a KS dictionary that is used to represent Kth-order tensor data. Tensor observations are assumed to be generated from a KS dictionary multiplied by sparse coefficient tensors that follow the separable sparsity model. This

work 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 errors $\{\varepsilon_k\}$ is shown to be $\max_{k\in[K]} \mathcal{O}(m_k p_k^3 \varepsilon_k^{-2})$.

7.1.1 Extensions of Sample Complexity Bounds

In terms of theoretical results, there are many aspects of KS-DL that have not been addressed in the literature so far. Firstly, our achievability result holds 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, we showed that there exists a gap between our provided sample complexity lower bounds and upper bounds. Hence, another future direction of possible interest includes using other techniques to find tighter bounds.

Moreover, the results that are obtained in Chapters 3 and 4 are based on the Frobenius norm distance metric and only provide local recovery guarantees. Open questions include corresponding abounds for other distance metrics and global recovery guarantees. In particular, getting global recovery guarantees requires using a distance metric that can handle the inherent permutation and sign ambiguities in the dictionary.

7.1.2 Algorithmic Open Problems

In terms of algorithmic open problems, a future direction of our work includes providing practical KS-DL algorithms that achieve the sample complexity scaling provided in Chapters 3 and 4.

Furthermore, in some cases we may not know a priori the parameters for which a KS dictionary yields a good model for the data. In particular, given dimension p, the problem of selecting the p_k's for coordinate dictionaries such that $p = \prod_k p_k$ has not been studied. For instance, in case of RGB images, selection of p_k's for the spatial modes is somewhat intuitive, as each column in the separable transform represents a

pattern in each mode. However, selecting the number of columns for the depth mode, which has 3 dimensions, is less obvious. This gives rise to the question: Given a fixed number of overall columns for the KS dictionary, how should we divide it between the number of columns for each coordinate dictionary?

7.2 Low Separation Rank Dictionary Learning for Tensor Data

In Chapter 5, we addressed the problem of learning sparse representations of tensor data using a mixture of separable dictionaries. For this purpose, we proposed the LSR-DL model to learn structured dictionaries. This model bridges the gap between unstructured and KS-DL models. We presented two LSR-DL algorithms called STARK and TeFDiL and showed that they have better generalization performance for image denoising in comparison to unstructured DL algorithm K-SVD [4] and existing KS-DL algorithms SeDiL [14] and BCD [16]. We also presented OSubDil that to the best our knowledge is the first online algorithm that results in LSR or KS dictionaries. We show that OSubDil results in a faster reduction in the reconstruction error in terms of number of observed samples compared to the state-of-the-art online DL algorithm [54] when the noise level in data is high.

7.2.1 Alternative Structures on Underlying Dictionary

In terms of future work, extensions of dictionary identifiability results to structures other than KS and LSR is an open problem. Examples of these structures include DL using the CP decomposition [120] and the tensor t-product [72]. Characterizing the DL problem and understanding the practical benefits of these models remain interesting questions for future work.

7.3 Massive MIMO Channel Estimation

In Chapter 6, we focused on multidimensional processing of tensor data and studied the advantages associated with taking the structure of data into account in tensor data representation. To this end, we studied the problem of sparse channel estimation in

(massive) MIMO-OFDM systems. We introduced the distinct block diagonal model for training data and obtained theoretical guarantees for channel recovery based on number of training pilot tones. Moreover, we studied the repetitive block diagonal model for training data that results in a Tucker decomposition for the observations. This formulation allows recovery of channel coefficients using sparse tensor recovery techniques that use less computational measures and memory compared to traditional recovery techniques. We further provided a comparison of the performance of the two models via numerical experiments. Future work in this regard includes providing formal guarantees for the repetitive block diagonal model.

7.3.1 Structured DL for Massive MIMO Channel Estimation

Future work in this regard includes the use of structured DL techniques to improve the channel estimation performance in massive MIMO-OFDM systems. Prior work by Ding and Rao [113] has shown that using a dictionary learning-based channel model for MIMO channel estimation results in an improved channel estimation performance compared to using predefined bases. This is attributed to the fact that the learned dictionary can adapt to the cell characteristics, can be applied to an arbitrary array geometry, and does not require accurate array calibration [113].

7.4 Joint Sparse Representations for Multimodal Data

Another interesting future direction of our work includes the problem of joint sparse representation of general multimodal data. Examples of such data include fMRI and EEG signals for a patient or a video with its corresponding audio signal. While various components of such multimodal data can be represented as tensors, the overall multimodal structure cannot be modeled as a multiway array. Although DL approaches that are based on tensor decompositions have resulted in efficient representations for tensor data, such methods cannot be extended to the problem of multimodal data representation. Prior works addressing this problem employ canonical correlation analysis (CCA) and its variants to boost cross-modal correlation for various modalities [121, 122] as well

as extensions of Latent Dirichlet Allocation (LDA) to model correlations of multimodal data at latent semantic (topic) level across modalities [123, 124]. Furthermore, Zhuang et al. [125] proposed a supervised coupled DL model with group structures for multimodal retrieval. This approach exploits the correlation between modalities using a set of linear mappings between sparse codes. Future work in this regard includes building upon these approaches to further exploit the tensor structure in the components as well as the overall multimodal structure in the data.

Bibliography

[1] 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, 2015.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal Component Analysis*. Springer, 1986, pp. 115–128.

[4] 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, 2006.

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

[6] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariance sparse coding for audio classification," in *Proc. 23rd Conf. Uncertainty in Artificial Intelligence*, 2007, pp. 149–158.

[7] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. learning*. ACM, 2007, pp. 759–766.

[8] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, 2012.

[9] Y. Rivenson and A. Stern, "Compressed imaging with a separable sensing operator," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 449–452, 2009.

[10] ——, "An efficient method for multi-dimensional compressive imaging," in *Frontiers in Optics 2009/Laser Science XXV/Fall 2009 OSA Optics & Photonics Technical Diges.* Optical Society of America, 2009, p. CTuA4.

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

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

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

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

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

[16] C. F. Caiafa and A. Cichocki, "Multidimensional compressed sensing and their applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 6, pp. 355–380, 2013.

[17] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, 2010.

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

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

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

[21] W. U. Bajwa, R. Calderbank, and D. G. Mixon, "Two are better than one: Fundamental parameters of frame coherence," *Appl. Comput. Harmon. Anal.*, vol. 33, no. 1, pp. 58–78, 2012.

[22] W. U. Bajwa and A. Pezeshki, "Finite frames for sparse signal processing," in *Finite Frames*, P. Casazza and G. Kutyniok, Eds. Cambridge, MA: Birkhäuser Boston, 2012, ch. 10, pp. 303–335.

[23] 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.

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

[25] S. Jokar and V. Mehrmann, "Sparse solutions to underdetermined Kronecker product systems," *Linear Algebra and its Appl.*, vol. 431, no. 12, pp. 2437–2447, 2009.

[26] R. A. Horn and C. R. Johnson, *Topics Matrix Anal.* Cambridge University Press, 1991.

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

[28] C. F. Caiafa and A. Cichocki, "Computing sparse representations of multidimensional signals using Kronecker bases," *Neural Comput.*, vol. 25, no. 1, pp. 186–220, 2013.

[29] R. N. Bracewell and R. N. Bracewell, *The Fourier transform and its applications*. McGraw-Hill New York, 1986, vol. 31999.

[30] I. Daubechies, *Ten lectures on wavelets*. SIAM, 1992, vol. 61.

[31] E. J. Candes and D. L. Donoho, "Curvelets: A surprisingly effective nonadaptive representation for objects with edges," *Curves and Surfaces*, pp. 105–120, 2000.

[32] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, 2005.

[33] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Human Genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[34] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[35] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006.

[36] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. Int. Conf. Artificial Neural Networks*. Springer, 1997, pp. 583–588.

[37] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[38] S. Li, J. Kawale, and Y. Fu, "Deep collaborative filtering via marginalized denoising auto-encoder," in *Proc. 24th ACM Int. Conf. Inf. and Knowledge Management*. ACM, 2015, pp. 811–820.

[39] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.

[40] J. M. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1395–1408, 2009.

[41] S. Mahdizadehaghdam, A. Panahi, H. Krim, and L. Dai, "Deep dictionary learning: A parametric network approach," *arXiv preprint arXiv:1803.04022*, 2018.

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

[43] R. Remi and K. Schnass, "Dictionary identification–sparse matrix-factorization via ℓ_1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[44] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *Conf. Learn. Theory*, 2012, pp. 37–1.

[45] Q. Geng and J. Wright, "On the local correctness of ℓ_1-minimization for dictionary learning," in *Proc. IEEE Int. Symp. Inf. Theory*. IEEE, 2014, pp. 3180–3184.

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

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

[48] O. Christensen, *An introduction to frames and Riesz bases*. Springer, 2016.

[49] K. A. Okoudjou, *Finite frame theory: a complete introduction to overcompleteness*. American Mathematical Soc., 2016, vol. 93.

[50] 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, 2014.

[51] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[52] V. Vapnik, "Principles of risk minimization for learning theory," in *Proc. Advances in Neural Inf. Process. Systems*, 1992, pp. 831–838.

[53] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 5. IEEE, 1999, pp. 2443–2446.

[54] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. Jan., pp. 19–60, 2010.

[55] H. Raja and W. U. Bajwa, "Cloud K-SVD: A collaborative dictionary learning algorithm for big, distributed data," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 173–188, 2016.

[56] Z. Shakeri, H. Raja, and W. U. Bajwa, "Dictionary learning based nonlinear classifier training from distributed data," in *Proc. 2nd IEEE Global Conf. Signal and Inf. Process.*, 2014, pp. 759–763.

[57] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images," *IEEE Trans. Image Process,*, vol. 21, no. 1, pp. 130–144, 2012.

[58] K. Schnass, "Local identification of overcomplete dictionaries," *J. Mach. Learn. Res.*, vol. 16, pp. 1211–1242, 2015.

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

[60] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity

recovery using ℓ_1-constrained quadratic programming (lasso)," *IEEE trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[61] P. Massart, *Concentration inequalities and model selection.* Springer, 2007, vol. 6.

[62] 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 Computational Advances in Multi-Sensor Adaptive Process.*, 2017, pp. 1–5.

[63] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Probl.*, vol. 27, no. 2, p. 025010, 2011.

[64] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[65] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds for Kronecker-structured dictionary learning," in *Proc. 2016 IEEE Int. Symp. Inf. Theory*, 2016, pp. 1148–1152.

[66] ——, "Sample complexity bounds for dictionary learning of tensor data," in *IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2017, pp. 4501–4505.

[67] ——, "Minimax lower bounds on dictionary learning for tensor data," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2706–2726, 2018.

[68] Z. Zhang and S. Aeron, "Denoising and completion of 3D data via multidimensional dictionary learning," in *Proc. 25th Int. Joint Conf. Artificial Intell.*, 2016, pp. 2371–2377.

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

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

[71] Y. Peng, D. Meng, Z. Xu, C. Gao, Y. Yang, and B. Zhang, "Decomposable nonlocal tensor dictionary learning for multispectral image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2949–2956.

[72] S. Soltani, M. E. Kilmer, and P. C. Hansen, "A tensor-based dictionary learning approach to tomographic image reconstruction," *BIT Numerical Math.*, pp. 1–30, 2015.

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

[74] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover, "Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 1, pp. 148–172, 2013.

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

[76] 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, 2017.

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

[78] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, 2009.

[79] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*. New York, NY USA: Cambridge University Press, 2009.

[80] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. John Wiley & Sons, 2012.

[81] J.-L. Durrieu, J. Thiran, F. Kelly *et al.*, "Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2012, pp. 4833–4836.

[82] J. von Neumann, "Some matrix inequalities and metrization of matrix space," *Tomsk Univ. Rev.*, vol. 1, no. 11, pp. 286–300, 1937, Reprinted in Collected Works (Pergamon Press, 1962), iv, 205 –219.

[83] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *Proc. 2010 IEEE Int. Symp. Inf. Theory*. IEEE, 2010, pp. 1373–1377.

[84] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[85] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Springer, 2013, vol. 1, no. 3.

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

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

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

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

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

[91] E. Schwab, B. Haeffele, N. Charon, and R. Vidal, "Separable dictionary learning with global optimality and applications to diffusion MRI," *arXiv preprint arXiv:1807.05595*, 2018.

[92] C. F. Dantas, J. E. Cohen, and R. Gribonval, "Learning fast dictionaries for sparse representations using low-rank tensor decompositions," in *Proc. Int. Conf. Latent Variable Analysis and Signal Separation.* Springer, 2018, pp. 456–466.

[93] K. Skretting and K. Engan, "Recursive least squares dictionary learning algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2121–2130, 2010.

[94] E. Dohmatob, A. Mensch, G. Varoquaux, and B. Thirion, "Learning brain regions via large-scale online structured sparse dictionary learning," in *Proc. Advances in Neural Inform. Process. Syst.*, 2016, pp. 4610–4618.

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

[96] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[97] K. Wimalawarne, M. Sugiyama, and R. Tomioka, "Multitask learning meets tensor factorization: Task imputation via convex optimization," in *Proc. Advances in Neural Inform. Process. Syst.*, 2014, pp. 2825–2833.

[98] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear

multitask learning," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28, no. 3, Atlanta, Georgia, USA, 2013, pp. 1444–1452.

[99] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Probl.*, vol. 27, no. 2, p. 025010, 2011.

[100] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[101] V. de Silva and L. Lim, "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1084–1127, 2008.

[102] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst. and Comput.*, vol. 1, 1993, pp. 40–44.

[103] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[104] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2010, pp. 2366–2369.

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

[106] Z. Shakeri, B. Taki, A. de Almeida, M. Ghassemi, and W. U. Bajwa, "Revisiting sparse channel estimation in massive MIMO-OFDM systems," in *Proc. Int Workshop Signal Process. Advances in Wireless Commun.* IEEE, 2019.

[107] Z. Shakeri and W. U. Bajwa, "Deterministic selection of pilot tones for compressive estimation of MIMO-OFDM channels," in *Proc. 48th Annu. Conf. Information Sciences and Systems*, 2015, pp. 1–6.

[108] M. Masood, L. H. Afify, and T. Y. Al-Naffouri, "Efficient coordinated recovery of sparse channels in massive mimo," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 104–118, 2015.

[109] W. Dongming, H. Bing, Z. Junhui, G. Xiqi, and Y. Xiaohu, "Channel estimation algorithms for broadband MIMO-OFDM sparse channel," in *14th IEEE Proc. Personal, Indoor and Mobile Radio Commun.*, vol. 2. IEEE, 2003, pp. 1929–1933.

[110] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in lte cellular networks: Key design issues and a survey," *IEEE Commun. Surv. Tutor.*, vol. 15, no. 2, pp. 678–700, 2013.

[111] D. C. Araújo, A. L. F. de Almeida, J. P. C. L. da Costa, and R. T. de Sousa, "Tensor-based channel estimation for massive MIMO-OFDM systems," *IEEE Access*, vol. 7, pp. 42 133–42 147, 2019.

[112] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[113] Y. Ding and B. D. Rao, "Dictionary learning-based sparse channel representation and estimation for FDD massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5437–5451, 2018.

[114] W. U. Bajwa, A. M. Sayeed, and R. Nowak, "A restricted isometry property for structurally-subsampled unitary matrices," in *Proc. Annu. Allerton Conf. Commun., Control, and Comput.*, 2009, pp. 1005–1012.

[115] W. U. Bajwa, "New information processing theory and methods for exploiting sparsity in wireless systems," Ph.D. dissertation, University of Wisconsin-Madison, Madison, WI, 2009.

[116] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[117] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure Appl. Math.*, vol. 61, no. 8, pp. 1025–1045, 2008.

[118] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, no. 2, pp. 489–509, 2006.

[119] A. Eftekhari, H. L. Yap, C. J. Rozell, and M. B. Wakin, "The restricted isometry property for random block diagonal matrices," *Appl. Comput. Harmon. Anal*, vol. 38, no. 1, pp. 1–31, 2015.

[120] Y. Zhang, X. Mou, G. Wang, and H. Yu, "Tensor-based dictionary learning for spectral CT reconstruction," *IEEE Trans. Med. Imaging*, vol. 36, no. 1, pp. 142–154, 2017.

[121] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[122] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*. ACM, 2010, pp. 251–260.

[123] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[124] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. 26th A. Int. ACM SIGIR Conf. Res. and Development in Inf. Retrieval*. ACM, 2003, pp. 127–134.

[125] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval." in *Proc. 27th AAAI Conf. Artificial Intell.*, 2013, pp. 1070–1076.