NONCONVEX MATRIX AND TENSOR RECOVERY WITH APPLICATIONS IN MACHINE LEARNING

BY MOHSEN GHASSEMI

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

Anand D. Sarwate

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2021

ABSTRACT OF THE DISSERTATION

Nonconvex Matrix and Tensor Recovery with Applications in Machine Learning

by Mohsen Ghassemi

Dissertation Director: Anand D. Sarwate

This thesis focuses on some fundamental problems in machine learning that are posed as nonconvex matrix factorizations. More specifically we investigate theoretical and algorithmic aspects of the following problems: i) inductive matrix completion (IMC), ii) structured dictionary learning (DL) from tensor data, iii) tensor linear regression and iv) principal component analysis (PCA). The theoretical contributions of this thesis include providing recovery guarantees for IMC and structured DL by characterizing the local minima and other geometric properties of these problems. The recovery results are stated in terms of upper bounds on the number of observations required to recover the true matrix (dictionary in the case of DL) underlying the data. Another major theoretical contribution of this work is providing fundamental limits on the performance of tensor linear regression solvers by deriving a lower bound on the worst case mean squared error of any estimator. On the algorithmic side, this thesis proposes novel online and batch algorithms for solving structured dictionary learning problem as well as a novel multi-stage accelerated stochastic PCA algorithm that achieves near optimal results.

Acknowledgements

I would like to express my deepest appreciation to my advisor, Prof. Anand Sarwate, for his support, patience, encouragement, and invaluable guidance. It has been a true pleasure working on many exciting and intellectually challenging problems under his guidance. He has taught me everything I know about how to approach and solve research problems, the value of mathematical rigor, and the art of presenting research results in an accessible way.

I would also like to extend my sincere thanks to my thesis committee members Prof. Waheed Bajwa, Prof. Mert Gurbuzbalaban, and Prof. Cunhui Zhang for reviewing my work and their invaluable insights and suggestions. I must especially thank Waheed who has been a mentor figure throughout my time at Rutgers and I have enjoyed collaborating with him on many research projects. I am also grateful to Mert for his guidance and collaboration on the work in Chapter 5 of this dissertation.

I was lucky to be surrounded with amazing friends and colleagues in Rutgers ECE departments who were always there for me and with whom I enjoyed many interesting discussion. I would specifically like to thank Zahra Shakeri, Haroon Raja, Talal Ahmed, Mohammad Hajimirsadeghi, and Saeed Soori; spending the past 7 years in Rutgers could not have been as much fun without them. Special thanks goes to my great friend and colleague Zahra with whom I enjoyed collaborating on the work in Chapter 3. I am also grateful to Baoul Taki who contributed to the work in Chapter 4. My thanks also goes out to my lab members Hafiz Imtiaz, Sijie Xiong, and Konstantinos Nikolakakis for their support and making my studies more enjoyable at Rutgers.

I have to thank my amazing parents and my brother Mohammad, without whom I would never be where I am today. I cannot thank them enough for their immense love and support. There is no way I can repay my parents for the countless sacrifices they

have made for me. Finally, my very special thanks go out to my soulmate and best friend Carolina. I am appreciative for her endless love and support. I would not have survived graduate school without the joy she, and our dog Lucy, have brought to my life.

Dedication

For my parents and my grandma.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Many fundamental problems in machine learning, statistics, and signal processing can be seen as *matrix estimation* problems. Examples of such problems include matrix sensing [1, 2], matrix completion [3–5], phase retrieval [6, 7], dictionary learning [8, 9], principal component analysis (PCA) [10], robust PCA [11], and blind deconvolution [12]. One common approach to solving these problems is resorting to convex relaxations and applying well-known convex optimization methods. While convexified formulations allow for employing well-established analytical tools to provide statistical performance guarantees for these problems, the computational cost and storage requirement of convex optimization methods makes them unsuitable for large-scale problems. Nonconvex matrix factorization schemes on the other hand enjoy lower storage requirements and per-iteration computational cost, and are amenable to parallelization. With prevalence of big data, these properties have become more important than ever since information processing and learning applications often involve dealing with high dimensional and/or high volume data, resulting in large-scale matrix factorization problems. Emergence of such large-scale problems necessitates development of efficient matrix factorization methods whose computational and storage costs scale favorably with matrix dimensions.

This thesis focuses on providing theoretical guarantees as well as developing efficient algorithms for some fundamental matrix factorization problems with applications in representation learning, recommendation systems, and other areas of machine learning.

1.2 Major Contributions

In this body of work we study three problems that can be formulated as nonconvex matrix decomposition problems. We first provide theoretical recovery guarantees for **inductive matrix completion (IMC)** by characterizing its optimization landscape. Then, we propose a novel **structured dictionary learning** model for learning sparse representations of tensor data and develop theory and numerical algorithms to validate the usefulness of this model. We also study the fundamental limits of estimation in a **tensor linear regression** problem and demonstrate the benefits of preserving the tensor structure of data and exploiting multi-directional interdependence among model variables in this problem. Finally, we develop a momentum-based accelerated algorithms for the streaming **principal component analysis (PCA)** problem and study the impact of introducing a momentum term in to a classic solver of this nonconvex problem. A more detailed overview is provided in the following.

In **Chapter 2**, we present our **first** major contribution. That is, we provide recovery guarantees for inductive matrix completion (IMC), a powerful technique with applications recommendation systems with side information. The aim of IMC is to reconstruct a low-rank matrix from a small number of given entries by exploiting the knowledge of the feature spaces of its row and column entities. We study the optimization landscape of this nonconvex problem and show that given sufficient number of observed entries, all local minima of the problem are globally optimum and all saddle points are "escapable". The immediate consequence of this result is that any first order optimization method such as gradient decent can be used to recover the true matrix. Moreover, we characterize how the knowledge of feature spaces reduces the number of required observed entries to recover (identify) the true matrix.

Our **second** main contribution, presented in **Chapter 3**, focuses on sparse representation learning for tensor (multidimensional) data. To this end, we study *dictionary learning* (DL), an effective and popular data-driven technique for obtaining sparse representations of data [8, 13], for tensor data. Traditional dictionary learning methods treat tensor data as vector data by collapsing each tensor to a vector. This disregards

the multidimensional structure in tensor data and results in dictionaries with large number of free parameters. With the increasing availability of large and high dimensional data sets, it is crucial to keep sparsifying models reasonably small to ensure their scalable learning and efficient storage. Our focus in this work is on learning of compact DL models that yield sparse representations of tensor data. Recently, some works have turned to tensor decompositions such as the Tucker decomposition [14] and CANDECOMP/PARAFAC decomposition (CPD) [15] for learning of *structured dictionaries* that have fewer number of free parameters. In particular, separable DL reduces the number of free parameters in the dictionary by assuming that the transformation on the sparse signal can be be implemented by performing a sequence of separate transformations on each signal dimension [16]. While separable DL methods enjoy lower sample/computational complexity and better storage efficiency over traditional DL [17] methods, the separability assumption among different modes of tensor data can be overly restrictive for many classes of data [18], resulting in an unfavorable trade-off between model compactness and representation power. In this work, we overcome this limitation by proposing a generalization of separable DL that we interchangeably refer to as learning a mixture of separable dictionaries or low separation rank DL (LSR-DL). This model provides better representation power than the separable model while having smaller number of parameters than traditional DL by allowing for increasing the number of parameters in structured DL in a consistent manner. To show the usefulness of our proposed model, we study the identifiability of the underlying dictionary in this model and derive the sufficient number of samples for local identification of the true dictionary under the LSR-DL model. Our results show that while the sample complexity of LSR-DL is slightly higher sample complexity than that of separable DL, it can still be significantly lower than that of traditional DL. We further develop efficient batch and online numerical algorithms to solve the LSR-DL problem.

Our **third** main contribution, which appears in **Chapter 4**, focuses on using an information theoretic approach to derive minimax risk (best achievable performance in the worst case scenario) of estimating the true parameter variables in a linear regression problem with tensor-structured data. Our results show a reduction in sample

complexity required for achieving a target worst case risk compared to the case where the data samples are treated as vectors and thus demonstrate the benefits of preserving the spatial structure of data and exploiting the multi-directional interdependence among model variables in the tensor linear regression model.

Finally, in **Chapter 5**, we present our **fourth** contribution. In this chapter we study the principal component analysis (PCA) problem when data arrives in a streaming fashion. We investigate the effect of adding a momentum term to the update rule of well-known stochastic PCA algorithm called Oja's method. While the efficacy of momentum-based acceleration for stochastic algorithms is not well-established in general, our proposed multi-stage accelerated variant of Oja's method achieves near optimal convergence rate in both in both noiseless case (bias term) and noisy case (variance term).

Chapter 2

Global Optimality in Inductive Matrix Completion

2.1 Introduction

Matrix completion [3, 19] is an important technique in machine learning with applications in areas such as recommendation systems [20] or computer vision [21]. The task is to reconstruct a low-rank matrix $\mathbf{M}^* \in \mathbb{R}^{n_1 \times n_2}$ from a small number of given entries. Theoretical results in the literature show that the number of required samples for exact recovery is $O(rn \log^2 n)$ where $n = n_1 + n_2$ and $r = \text{rank}(\mathbf{M}^*)$ [22, 23]. In some applications, the algorithm may have access to *side information* that can be exploited to improve this sample complexity. For example, in many recommendation systems where the given entries of $\mathbf{M}^*$ represent the ratings given by users (row entities) to items (column entities), the system has additional information about both user profiles and items.

Among the many approaches to incorporate side information [24–30], *inductive matrix completion (IMC)* [24, 25] models side information as knowledge of *feature spaces*. This is incorporated in the model by assuming that each entry of the unknown matrix of interest $\mathbf{M}^* \in \mathbb{R}^{n_1 \times n_2}$ is in form of $\mathbf{M}^*_{ij} = \mathbf{x}_i^T \mathbf{W}^* \mathbf{y}_j$, where $\mathbf{x}_i \in \mathbb{R}^{d_1}$ and $\mathbf{y}_j \in \mathbb{R}^{d_2}$ are known feature vectors of i-th row (user) and j-th column (item), respectively. The low-rank matrix completion problem in this case can be formulated as recovering a rank-r matrix $\mathbf{W}^* \in \mathbb{R}^{d_1 \times d_2}$ such that the observed entries are $\mathbf{M}_{ij} = \mathbf{x}_i^T \mathbf{W}^* \mathbf{y}_j$. In fact, the IMC problem translates to finding missing entries of $\mathbf{M}^*$ as recovering matrix $\mathbf{W}^*$ from its measurements in form of $\mathbf{x}_i \mathbf{W}^* \mathbf{y}_j = \left\langle \mathbf{x}_i \mathbf{y}_j^T, \mathbf{W}^* \right\rangle$ for $(i, j) \in \Omega$.

Using this model, the sample complexity decreases considerably if the size of matrix $\mathbf{M}$ is much larger than $\mathbf{W}^*$. Another advantage of this model is that rows/columns of the unknown matrix can be predicted without knowing even one of their entries using

the corresponding feature vectors once we recover $\mathbf{W}^*$ using the given entries. This is not possible in standard matrix completion since a necessary condition for completing a rank-r matrix is that at least r entries of every row and every column are observed [3].

The nonconvex rank-r constraint makes the problem challenging. There are two main approaches in the matrix recovery literature to impose the low-rank structure in a tractable way. The first approach is using convex relaxations of the nonconvex rank-constrained problem [3, 11, 23, 31–33]. In the IMC problem, at least $O(rd \log d \log n)$ samples are required for recovery of $\mathbf{W}^*$ using a trace-norm relaxation, where $d = d_1 + d_2$ [24, 25]. The trace-norm approach has also been proposed for the IMC problem with noisy features where the unknown matrix is modeled as $\mathbf{X}\mathbf{W}^*\mathbf{Y}^T + \mathcal{N}$ where the residual matrix N models imperfections and noise in the features [27].

Another approach uses matrix factorization, where the $d_1 \times d_2$ matrix $\mathbf{W}$ is expressed as $\mathbf{W} = \mathbf{U}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ [20, 34]. Jain et al. show that alternating minimization (AM) converges to the global solution of matrix sensing and matrix completion problems in linear time under standard conditions [34]. Inspired by this result, Zhong et al. [25] show that for the factorized IMC problem, $O(r^3 d \log d \max\{r, \log n\})$ samples are sufficient for ϵ-recovery of $\mathbf{W}^*$ using AM.

On the computational side, the per-iteration cost of the solvers of the convex matrix estimation problem is high since they require finding the SVD of a matrix in case of implementing singular value thresholding [35] or proximal gradient methods [36], or they involve solving a semi-definite program. On the other hand, both empirically and theoretically, stochastic gradient descent (SGD) and AM have been shown to find good local optima in many *nonconvex* matrix estimation problems and that suitable modifications to the objective function can find *global optima* [34, 37]. These simple local search algorithms have low memory requirement and per-iteration computational cost, due to the fact that in low-rank problems $r \ll d_1, d_2$. Although the IMC model reduces the dimensionality of the matrix estimation problem from $n_1 \times n_2$ to $d_1 \times d_2$, the lower complexity of the solvers of the factorized model is appealing [25].

On the theoretical side, the trace-norm based model is intriguing in that it allows for employing well-established tools to analyze the statistical performance of the convex

program. Although the matrix factorization based models in general are theoretically less understood, recent works have studied the optimization landscape of some of these nonconvex problems and show that their objective functions are devoid of "poor" local minima. Problems such as matrix completion [4, 5], matrix sensing [1, 2], phase retrieval [7], deep (linear) neural networks [38, 39] are amenable to this approach. To the best of our knowledge, this work is the first to study the geometry and the statistical performance of IMC under the factorized model.

The work in this chapter is motivated by the recovery guarantees of AM for the (nonconvex) factorized IMC problem. Our key technical contribution is to use concentration inequalities to show that given a sufficient number of measurements, the ensemble of sensing matrices $\mathbf{x}_i\mathbf{y}_j^T$ almost preserves the energy of all rank-$2r$ matrices, i.e. it satisfies *restricted isometry property* of order $2r$. This allows us to use the framework of Ge et al. [5] for matrix sensing problems. Our final result is that given at least $O(dr\max\{r^2, \log^2 n\})$ observations, in the (regularized) factorized IMC problem *i*) all local minima are globally optimal, *ii*) all local minima fulfill $\mathbf{U}\mathbf{V}^T = \mathbf{W}^*$, and *iii*) the saddle points are escapable in the sense that the Hessian at those points has at least one negative eigenvalue.

Our result implies that the success of AM in the nonconvex IMC problem is to some degree a result of the geometry of the problem and not solely due to the properties of the algorithm. In fact, any algorithm with guaranteed convergence to a local minimum, e.g. SGD [37], can be used for solving the factorized IMC problem.[1]

2.2 Problem Model

Notation and Definitions. Throughout this chapter, vectors and matrices are, respectively, denoted by boldface lower case letters: $\mathbf{a}$ and boldface upper case letters: $\mathbf{A}$. We denote by $\mathbf{A}_{ij}$ the j-th element of the i-th row of $\mathbf{A}$. The smallest eigenvalue of $\mathbf{A}$ is denoted by $\lambda_{\min}(\mathbf{A})$. In matrix completion, the set of indices of the observed (given) entries of an incomplete matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ is denoted by Ω with size $m = |\Omega|$. Also,

[1] The results presented in this chapter have been published in Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing [40]

$\mathbf{A}_\Omega$ denotes the linear projection of $\mathbf{A}$ onto the space of $n_1 \times n_2$ matrices whose entries outside Ω are zero. The inner product of two matrices is defined as $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^T\mathbf{B})$. In a noncovex optimization problem, a poor local minimum is a local minimum which is not globally optimum.

We repeatedly use the *(matrix) restricted isometry property (RIP)* [31] and the *strict saddle property* [37, 41] defined below.

Definition 1. *A linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ satisfies r-RIP with δ_r RIP constant if for every $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ such that $rank(W) \leq r$ it holds that*

$$(1 - \delta_r) \|\mathbf{W}\|_F^2 \leq \|\mathcal{A}(\mathbf{W})\|_2^2 \leq (1 + \delta_r) \|\mathbf{W}\|_F^2 .$$

Definition 2. *A twice differentiable function $f(\mathbf{x})$ is strict saddle if $\lambda_{\min}\left(\nabla^2 f(x)\right) < 0$ at its saddle points.*

Inductive Matrix Completion. Consider the nonconvex low-rank matrix completion problem

$$\min_{\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}} \|\mathbf{M}_\Omega^* - \mathbf{M}_\Omega\|_F^2 \quad \text{s.t.} \quad \mathrm{rank}(\mathbf{M}) \leq r. \tag{2.1}$$

In an inductive matrix completion problem, the underlying matrix has the form $\mathbf{M}^* = \mathbf{X}\mathbf{W}^*\mathbf{Y}^T$ where the *side information* matrices $\mathbf{X} \in \mathbb{R}^{n_1 \times d_1}$ and $\mathbf{Y} \in \mathbb{R}^{n_2 \times d_2}$ are known and $\mathbf{W}^* = \mathbf{U}^*\mathbf{V}^{*T}$ with $\mathbf{U}^* \in \mathbb{R}^{d_1 \times r},\ \mathbf{V}^* \in \mathbb{R}^{d_2 \times r}$ is unknown. Therefore, the problem can be written as

$$\begin{aligned} &\min_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} \left\| \left(\mathbf{M}^* - \mathbf{X}\mathbf{W}\mathbf{Y}^T\right)_\Omega \right\|_F^2 \\ &\text{s.t.} \quad \mathrm{rank}(\mathbf{W}) \leq r. \end{aligned} \tag{2.2}$$

This problem can be reformulated into an unconstrained nonconvex problem by expressing $\mathbf{W}$ as $\mathbf{U}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{d_1 \times r},\ \mathbf{V} \in \mathbb{R}^{d_2 \times r}$:

$$\min_{\mathbf{U},\mathbf{V}} \ f(\mathbf{U}, \mathbf{V}) = \left\| \left(\mathbf{M}^* - \mathbf{X}\mathbf{U}\mathbf{V}^T\mathbf{Y}^T\right)_\Omega \right\|_F^2 + R(\mathbf{U}, \mathbf{V}) \tag{2.3}$$

The regularization term $R(\mathbf{U}, \mathbf{V})$ is added to account for the invariance of the asymmetric factorized model to scaling of the factor matrices by reciprocal values. A common choice that suits our model is $R(\mathbf{U}, \mathbf{V}) = \frac{1}{4}\|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\|_F^2$ [2, 5].

The objective function $f(\mathbf{U}, \mathbf{V})$ in problem (2.3) alternatively can be written as

$$f(\mathbf{U}, \mathbf{V}) = \sum_{(i,j)\in\Omega} \left(\mathbf{M}^*_{ij} - \left\langle \mathbf{x}_i \mathbf{y}_j^T, \mathbf{U}\mathbf{V}^T \right\rangle\right)^2 + \frac{1}{4}\left\|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\right\|_F^2,$$

where $\mathbf{x}_i^T$ and $\mathbf{y}_j^T$ respectively are the ith and jth rows of $\mathbf{X}$ and $\mathbf{Y}$. Observe that $\left\langle \mathbf{x}_i \mathbf{y}_j^T, \mathbf{U}\mathbf{V}^T \right\rangle = \mathbf{x}_i^T \mathbf{U}\mathbf{V}^T \mathbf{y}_j$.

This shows that the IMC problem (2.3) can be thought of as a matrix sensing problem where we are given linear measurements of the $d_1 \times d_2$ matrix $\mathbf{W}^*$ by sensing matrices $\mathbf{A}_{ij} = \mathbf{x}_i \mathbf{y}_j^T$. Define the linear operator $\mathcal{A}$ such that $\mathcal{A}(\mathbf{W})$ is a vector whose elements are the measurements $\frac{1}{\sqrt{m}} \langle \mathbf{A}_{ij}, \mathbf{W} \rangle$.

In this chapter, we make the following assumptions regarding the side information matrices and the sampling model.

Assumption 1 (Side information)**.** *The side information matrices satisfy* $\mathbf{X}^T\mathbf{X} = n_1 I_{d_1}$ *and* $\mathbf{Y}^T\mathbf{Y} = n_2 I_{d_2}$.[2] *We also make the assumption that for any given matrices* $\bar{\mathbf{U}}$ *and* $\bar{\mathbf{V}}$ *with orthogonal columns, the rows of the side information matrices (feature vectors) satisfy* $\left\|\bar{\mathbf{U}}\mathbf{x}_i\right\|_2^2 \leq \mu\bar{r}$ *and* $\left\|\bar{\mathbf{V}}\mathbf{y}_j\right\|_2^2 \leq \mu\bar{r}$*, where* $\bar{r} = \max(r, \log n_1, \log n_2)$ *and* μ *is a positive constant. This assumption, for example, is satisfied with high probability when the side information matrices* $\mathbf{X}$ *and* $\mathbf{Y}$ *are instances generated from a random orthonormal matrix model (the first* d_1 *(respectively* d_2*) columns) and rescaled by* $\sqrt{n}_1$ *(respectively* $\sqrt{n}_2$*) [3, 25].*

Assumption 2 (Sampling model)**.** *Indices* i *and* j *are independent and uniformly distributed on* $\{1, 2, \ldots, n\}$.

[2]This is not a restrictive assumption since we can apply orthonormalization methods such as Gram-Schmidt process [42] and then rescale to ensure this assumption is satisfied.

2.3 Geometric Analysis

We are interested in the geometric landscape of the objective function in the IMC problem (2.3). We will show that simple algorithms like AM can recover the true underlying matrix with arbitrary accuracy because given enough observations, the objective function in this problem *i*) has no poor local minima, *ii*) has only local minima which satisfy $\mathbf{U}\mathbf{V}^T = \mathbf{W}^*$, and *iii*) is *strict saddle*.

We employ the framework developed by Ge et al. [5] for matrix sensing to show that the objective function of the IMC problem (2.3) satisfies properties *i*), *ii*), and *iii*). Theorem 1 states the main result of this chapter.

Theorem 1. *Consider the IMC problem* (2.3) *seen as a matrix recovery problem with sensing matrices* $\mathbf{A}_{ij} = \mathbf{x}_i\mathbf{y}_j^T$ *for* $(i,j) \in \Omega$*, such that Assumptions 1 and 2 hold. Let* $\bar{r} = \max\{r, \log n_1, \log n_2\}$*. If the number of measurements is* $m = O\left(\mu^2 dr^2\bar{r}\right)$*, then there exists a positive constant* h *such that with probability higher than* $1-2\exp(-hm)$*, the nonconvex objective function* $f(\mathbf{U},\mathbf{V})$ *has the following properties: i) all its local minima are globally optimal , ii) all its local minima satisfy* $\mathbf{U}\mathbf{V}^T = \mathbf{M}^*$*, and iii) it satisfies the strict saddle property.*

The proof strategy here is to show that at any stationary point of $f(\mathbf{U},\mathbf{V})$ (and its neighborhood), the "difference" $\mathbf{\Delta}$ between the point and the true solution (which is basically the Euclidian distance between the point and its nearest global minimum) is a *descent direction*. This means that $(\mathbf{U},\mathbf{V})$ cannot be local minimum unless $\mathbf{\Delta} = \mathbf{0}$ (no poor local minima and exact recovery) and that the Hessian at the saddle points cannot be positive semidefinite (strict saddle property). To this end, following the proposed strategy by Ge, Jin, and Zheng [5], we construct $\mathbf{B} = \begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix} \in \mathbb{R}^{(d_1+d_2)\times r}$, $\mathbf{W} = \mathbf{U}\mathbf{V}^T$, and $\mathcal{N} = \mathbf{B}\mathbf{B}^T$ and reformulate problem (2.3) as the positive semidefinite (PSD) low-rank matrix recovery problem

$$\min_{\mathbf{B}} f(\mathbf{B}) = \|\mathcal{T}(\mathcal{N}^*) - \mathcal{T}(\mathbf{B}\mathbf{B}^T)\|_2^2. \tag{2.4}$$

where $\mathbf{B}^* = \begin{pmatrix} \mathbf{U}^* \\ \mathbf{V}^* \end{pmatrix}$, $\mathcal{N}^* = \mathbf{B}^*\mathbf{B}^{*T}$, and $\mathcal{T}$ is a linear operator such that $\mathcal{T}(\mathcal{N})$ is an ensemble of m measurements $\langle \mathbf{T}_{ij}, \mathcal{N} \rangle$ such that

$$\langle \mathbf{T}_{ij}, \mathcal{N} \rangle^2 = \frac{1}{m} \left(4 \langle \mathbf{A}_{ij}, \mathbf{W} \rangle^2 + \| \mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T \|_F^2 \right).$$

The following definition captures the invariance of the solution of symmetric matrix recovery to rotation, negation, or column permutation.

Definition 3. *Given matrices* $\mathbf{B}, \mathbf{B}^* \in \mathbb{R}^{d \times r}$*, define the rotation invariant difference* $\boldsymbol{\Delta}(\mathbf{B}; \mathbf{B}*) \triangleq \mathbf{B} - \mathbf{B}^*\mathbf{D}$*, where* $\mathbf{D} = \underset{\mathbf{Z}:\mathbf{Z}\mathbf{Z}^T = \mathbf{Z}^T\mathbf{Z} = \mathbf{I}_r}{\operatorname{argmin}} \| \mathbf{B} - \mathbf{B}^*\mathbf{Z} \|_F^2$.

We use the shorthand $\boldsymbol{\Delta}$ for $\boldsymbol{\Delta}(\mathbf{B}; \mathbf{B}^*)$ in the rest of this chapter. The second order term in the Taylor expansion of $f(\mathbf{B})$ becomes dominant in the neighborhood of stationary points. Therefore it suffices to show that $\boldsymbol{\delta}^T \nabla^2 f(\mathbf{B}) \boldsymbol{\delta}$, where $\boldsymbol{\delta} = \operatorname{vec}(\boldsymbol{\Delta})$, is strictly negative for points in these regions, except when $\boldsymbol{\Delta} = \mathbf{0}$, to prove that $\boldsymbol{\Delta}$ is a descent direction. Theorem 2 states that if linear operator $\mathcal{B}$ is RIP, then we can show $\boldsymbol{\delta}^T \nabla^2 f(\mathbf{B}) \boldsymbol{\delta}$ is strictly negative in the neighborhood of stationary points unless they correspond to $\mathcal{N}^*$ (and its submatrix $\mathbf{W}^*$) and consequently $\mathbf{M}^* = \mathbf{X}\mathbf{W}^*\mathbf{Y}^T$, the ground truth matrix in problem (2.3).

Theorem 2. *Consider the objective function of the PSD matrix recovery problem* (2.4)*. If the measurement operator* $\mathcal{T}$ *satisfies* $(2r, \delta_{2r})$*-RIP, then any point satisfying* $\| \nabla f(\mathbf{B}) \|_F \leq \xi$*, the quadratic form* $\boldsymbol{\delta}^T \nabla^2 f(\mathbf{B}) \boldsymbol{\delta}$ *for* $\boldsymbol{\delta} = \operatorname{vec}(\boldsymbol{\Delta})$ *defined above is negative unless* $\| \boldsymbol{\Delta} \|_F \leq K\xi / (1 - 5\delta_{2r})$ *for some positive constant* K*.*

Proof sketch. The proof is based on the following equality (Lemma 7 in [5]):

$$\boldsymbol{\delta}^T \nabla^2 f(\mathbf{B}) \boldsymbol{\delta} = \left\| \mathcal{T}\left(\boldsymbol{\Delta}\boldsymbol{\Delta}^T \right) \right\|_2^2 - 3 \left\| \mathcal{T}\left(\mathcal{N} - \mathcal{N}^* \right) \right\|_2^2 + 4 \left\langle \nabla f(\mathbf{B}), \boldsymbol{\Delta} \right\rangle. \tag{2.5}$$

Using the RIP property of $\mathcal{T}$, which implies that the measuring operator captures the energy of the observed matrix with small deviation, and applying the bounds

$\left\|\boldsymbol{\Delta}\boldsymbol{\Delta}^T\right\|_F^2 \leq 2\left\|\mathcal{N}-\mathcal{N}^*\right\|_F^2$ and $k\left\|\boldsymbol{\Delta}\right\|_F^2 \leq \left\|\mathcal{N}-\mathcal{N}^*\right\|_F^2$ (Lemma 6 in [5]) results in

$$\boldsymbol{\delta}^T\nabla^2 f(\mathbf{B})\boldsymbol{\delta} \leq -k\left(1-5\delta_{2r}\right)\left\|\boldsymbol{\Delta}\right\|_F^2 + 4\xi\left\|\boldsymbol{\Delta}\right\|_F. \tag{2.6}$$

Therefore the bilinear form on the left cannot be nonnegative unless we have $\left\|\boldsymbol{\Delta}\right\|_F^2 \leq 4\xi/\left(k\left(1-5\delta_{2r}\right)\right)$.

□

Now, we show that the linear operator $\mathcal{A}$ and consequently $\mathcal{T}$ are $2r$-RIP. Note that it is important that we show $2r$-RIP rather that r-RIP because in Theorem 2, $\mathcal{T}$ is applied to $\mathbf{B}-\mathbf{B}^*$ which can be of rank at most $2r$. It also guarantees that the null space of $\mathcal{T}$ does not include any matrices of rank $2r$ or less, which is a necessary and sufficient condition for unique recovery [43, 44].

Theorem 3. *Consider the IMC problem* (2.3) *seen as a matrix recovery problem with sensing matrices* $\mathbf{A}_{ij} = \mathbf{x}_i\mathbf{y}_j^T$ *for* $(i,j) \in \Omega$*, such that Assumptions 1 and 2 hold. If the number of measurements* $m = O\left(\mu^2 d\bar{r}^2 r\log(36\sqrt{2}/\delta)/\delta^2\right)$*, then there exists a positive constant* h *such that with probability higher than* $1-2\exp\left(-hm\right)$*, the linear operator* $\mathcal{A}(\cdot)$*, seen as an ensemble of* m *measurements* $\frac{1}{\sqrt{m}}\left\langle\mathbf{A}_{ij},\cdot\right\rangle$*, is* $2r$*-RIP with RIP constant* $\delta_{2r} = 2\delta$*.*

Proof. We show that $\left\|\mathcal{A}(\widetilde{\mathbf{W}})\right\|_2^2$ is close to $\left\|\widetilde{\mathbf{W}}\right\|_F^2$ for all rank-$2r$ matrices $\widetilde{\mathbf{W}}$, i.e., $\left|\left\|\mathcal{A}(\widetilde{\mathbf{W}})\right\|_2^2 - \left\|\widetilde{\mathbf{W}}\right\|_F^2\right| \leq \delta_{2r}\left\|\widetilde{\mathbf{W}}\right\|_F^2$. We use Bernstein's inequality to find a bound on the deviation of the sum of m random variables $\frac{1}{\sqrt{m}}\left\langle\mathbf{x}_i\mathbf{y}_j^T,\widetilde{\mathbf{W}}\right\rangle$ from their mean $\left\|\widetilde{\mathbf{W}}\right\|_F^2$ for a given rank-$2r$ matrix $\widetilde{\mathbf{W}}$. This is formally stated in Lemma 1. Then we find a similar bound for all rank-$2r$ (or less) matrices.

Lemma 1. *Consider the same setting as Theorem 3. For a given matrix* $\widetilde{\mathbf{W}}$ *of rank* $2r$*, with probability at least* $1-C\exp\left(-cm\right)$*, for some positive constatnts* C *and* c*, we have*

$$\left(1-\delta_{2r}\right)\left\|\widetilde{\mathbf{W}}\right\|_F^2 \leq \left\|\mathcal{A}\left(\widetilde{\mathbf{W}}\right)\right\|_2^2 \leq \left(1+\delta_{2r}\right)\left\|\widetilde{\mathbf{W}}\right\|_F^2.$$

Proof of Lemma 1. In order to show that the average random measurement, denoted by $\|\mathcal{A}(\widetilde{\mathbf{W}})\|_2^2 = \frac{1}{m}\sum_{ij}\langle \mathbf{A}_{ij}, \widetilde{\mathbf{W}}\rangle^2$, is close to its expectation $\|\widetilde{\mathbf{W}}\|_F^2$, we use Bernstein's inequality [45]:

$$\mathbb{P}\left(\left|\bar{Z} - \eta_Z\right| > \epsilon\right) \leq 2\exp\Big(\frac{-m\epsilon^2/2}{\frac{1}{m}\sum_{ij}\mathrm{Var}(Z_{ij}) + B_Z\epsilon/3}\Big),$$

where $\bar{Z} = \frac{1}{m}\sum_{ij} Z_{ij}$ and η_Z is the mean of the random variables. To apply Bernstein's inequality, we need to find the expectation, the variance (or an upper bound on the variance), and an upper bound on the absolute value of the random variables in the summand, denoted by $Z_{ij} = \mathbf{x}_i^T\widetilde{\mathbf{W}}\mathbf{y}_j\mathbf{y}_j^T\widetilde{\mathbf{W}}\mathbf{x}_i$. Note that $\mathbf{X}$ and $\mathbf{Y}$ are known orthogonal matrices and the only source of randomness is the choice of (i,j). First, we find the mean of the random variables:

$$\begin{aligned}
\eta_Z &= \mathbb{E}\left[\mathbf{x}_i^T\widetilde{\mathbf{W}}\mathbf{y}_j\mathbf{y}_j^T\widetilde{\mathbf{W}}\mathbf{x}_i\right] \\
&= \mathbb{E}\left[e_i^T\mathbf{X}\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T\mathbf{Y}^T e_j e_j^T\mathbf{Y}\widetilde{\mathbf{W}}^T\mathbf{X}^T e_i\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\widetilde{\mathbf{V}}^T\mathbf{Y}^T e_j e_j^T\mathbf{Y}\widetilde{\mathbf{W}}^T\mathbf{X}^T e_i e_i^T\mathbf{X}\widetilde{\mathbf{U}}\right)\right] \\
&= \mathrm{Tr}\left(\widetilde{\mathbf{V}}^T\mathbf{Y}^T\mathbb{E}\left[e_j e_j^T\right]\mathbf{Y}\widetilde{\mathbf{W}}^T\mathbf{X}^T\mathbb{E}\left[e_i e_i^T\right]\mathbf{X}\widetilde{\mathbf{U}}\right) \\
&\overset{(a)}{=} \mathrm{Tr}\left(\widetilde{\mathbf{V}}^T\mathbf{Y}^T\mathbf{Y}\widetilde{\mathbf{W}}^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{U}}\right) \\
&\overset{(b)}{=} \mathrm{Tr}\left(\widetilde{\mathbf{V}}^T\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}\right) \\
&= \mathrm{Tr}\left(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T\cdot\widetilde{\mathbf{W}}^T\right) \\
&= \|\widetilde{\mathbf{W}}\|_F^2,
\end{aligned} \tag{2.7}$$

where $\widetilde{\mathbf{W}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T$, equality (a) follows from $\mathbb{E}\left[e_i e_i^T\right] = \frac{1}{n_1}I_{n_1}$ and (b) follows from Assumption 1. Next we find an upper bound B_Z on $|Z_{ij}|$:

$$\begin{aligned}
|Z_{ij}| &= \cdot\mathbf{x}_i^T\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^T\mathbf{y}_j\cdot\mathbf{y}_j^T\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^T\mathbf{x}_i \\
&\leq \left(\left\|\mathbf{x}_i^T\widehat{\mathbf{U}}\right\|_2\left\|\widehat{\mathbf{\Sigma}}\right\|_2\left\|\widehat{\mathbf{V}}^T\mathbf{y}_j\right\|_2\right)^2 \\
&= \sigma_1^2\left\|\widehat{\mathbf{V}}^T\mathbf{y}_j\right\|_2^2\cdot\left\|\widehat{\mathbf{U}}^T\mathbf{x}_i\right\|_2^2 \\
&\leq \bar{r}^2\mu^2\sigma_1^2,
\end{aligned} \tag{2.8}$$

where $\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^T$ is the SVD of $\widetilde{\mathbf{W}}$, $\sigma_1 = \|\widetilde{\mathbf{W}}\|_2$, and the last inequality follows from Assumption 1. Finally, for the variance of the random variables we have

$$\begin{aligned}\frac{1}{m}\sum_{ij}\mathrm{Var}(Z_{ij}) &\leq \frac{1}{m}\sum_{ij}\mathbb{E}\left[Z_{ij}^2\right] \\ &\overset{(a)}{\leq} \frac{1}{m}B_z\sum_{ij}\mathbb{E}\left[Z_{ij}\right] \\ &\leq \bar{r}^2\mu^2\sigma_1^2\|\widetilde{\mathbf{W}}\|_F^2, \end{aligned} \tag{2.9}$$

where inequality (a) is due to the fact that Z_{ij}'s are nonnegative random variables. Using Bernstein's inequality we get the following.

$$\begin{aligned}\mathbb{P}\Big(\Big|\|\mathcal{A}(\widetilde{\mathbf{W}})\|_2^2 - \|\widetilde{\mathbf{W}}\|_F^2\Big| > \epsilon\Big) &\leq 2\exp\left(-\frac{m\epsilon^2/2}{\frac{1}{m}\sum_{ij}\mathrm{Var}(Z_{ij}) + B_Z\epsilon/3}\right) \\ &\leq 2\exp\left(-\frac{m\epsilon^2/2}{\bar{r}^2\mu^2\sigma_1^2\|\widetilde{\mathbf{W}}\|_F^2 + \bar{r}^2\mu^2\sigma_1^2\epsilon/3}\right). \end{aligned} \tag{2.10}$$

Set $\epsilon = \delta\,\|\mathbf{W}\|_F^2$. We have

$$\begin{aligned}\mathbb{P}\Big(\Big|\|\mathcal{A}(\widetilde{\mathbf{W}})\|_2^2 - \|\widetilde{\mathbf{W}}\|_F^2\Big| > \delta\|\widetilde{\mathbf{W}}\|_F^2\Big) &\leq 2\exp\left(-\frac{m\delta^2\|\widetilde{\mathbf{W}}\|_F^2/2}{\bar{r}^2\mu^2\sigma_1^2(1+\delta/3)}\right) \\ &\leq 2\exp\left(-\frac{m\delta^2/2}{\mu^2\bar{r}^2(1+\delta/3)}\right). \end{aligned} \tag{2.11}$$

Set $\delta = \sqrt{\frac{4\mu^2\bar{r}^2\log(2/\rho)}{m}}$. If $m > 4\mu^2\bar{r}^2\log(2/\rho)$ we have $\delta < 1$. Therefore,

$$\mathbb{P}\Big(\Big|\|\mathcal{A}(\widetilde{\mathbf{W}})\|_2^2 - \|\widetilde{\mathbf{W}}\|_F^2\Big| > \delta\|\widetilde{\mathbf{W}}\|_F^2\Big) \leq 2\exp\left(-\frac{m\delta^2}{4\mu^2\bar{r}^2}\right).$$

This concludes the proof of Lemma 1. □

Now we return to the proof of Theorem 3. The rest of the proof is based on Theorem 2.3 in [46]. We showed in Lemma 1 that for a given matrix of rank at most $2r$,

$$\mathbb{P}\Big(\Big|\|\mathcal{A}(\widetilde{\mathbf{W}})\|_2^2 - \|\widetilde{\mathbf{W}}\|_F^2\Big| > \delta\|\widetilde{\mathbf{W}}\|_F^2\Big) \leq C\exp(cm),$$

for positive constants C and c. In order to extend the result such that a similar result holds for all rank-$2r$ (or less) matrices, we use the union bound for an ϵ-net [3] [47] of the space of such matrices with unit Frobenius norm. For the set $\mathbb{S}_{2r}^d = \{\widetilde{\mathbf{W}} \in \mathbb{R}^{d\times d} : \ \text{rank}(\widetilde{\mathbf{W}}) \leq 2r, \left\|\widetilde{\mathbf{W}}\right\|_F = 1\}$, there exists an ϵ'-net $\bar{\mathbb{S}}_{2r}^d \subset \mathbb{S}_{2r}^d$ such that $|\bar{\mathbb{S}}_{2r}^d| \leq (9/\epsilon')^{(2d+1)2r}$ [43, 46]. It follows from 2.12 and the union bound that

$$\mathbb{P}\Big(\max_{\widetilde{\widetilde{\mathbf{W}}}\in\bar{\mathbb{S}}_{2r}^d} \Big| \left\|\mathcal{A}\left(\widetilde{\widetilde{\mathbf{W}}}\right)\right\|_2^2 - 1\Big| > \delta\Big) \leq |\hat{S}_{2r}^d| C\exp(-cm).$$

Setting $\epsilon' = \delta/(4\sqrt{2})$ results in

$$\begin{aligned}\mathbb{P}\Big(\max_{\widetilde{\widetilde{\mathbf{W}}}\in\bar{\mathbb{S}}_{2r}^d} \Big| \left\|\mathcal{A}\left(\widetilde{\widetilde{\mathbf{W}}}\right)\right\|_2^2 - 1\Big| > \delta\Big) &\leq C\exp\Big((2d+1)2r\log(36\sqrt{2}/\delta) - cm\Big) \\ &= C\exp\big(c'dr - cm\big) \\ &\leq C\exp\left(-hm\right), \end{aligned} \tag{2.12}$$

where $c' = 6\log(36\sqrt{2}/\delta)$ and $h = c - c'/(K)$. We need $m > Kdr$ so that the last inequality above holds, and we need $K > c'/c$ so that h becomes positive. This means that $m > c'dr/c$. Plugging in the values for C, c, and c', we get that if with probability at least $1 - 2\exp\left(-hm\right)$,

$$\max_{\widetilde{\widetilde{\mathbf{W}}}\in\bar{\mathbb{S}}_{2r}^d} \left| \left\|\mathcal{A}(\widetilde{\widetilde{\mathbf{W}}})\right\|_2^2 - 1\right| \leq \delta.$$

It follows from this bound that for all $\widetilde{\mathbf{W}}$ of rank at most $2r$ that with probability at least $1 - 2\exp\left(-hm\right)$ [46],

$$1 - 2\delta \leq \left\|\mathcal{A}\Big(\frac{\widetilde{\mathbf{W}}}{\|\widetilde{\mathbf{W}}\|_F^2}\Big)\right\|_2^2 \leq 1 + 2\delta.$$

[3] An ϵ-net is a set of points such that the union of radius-ϵ balls centered at these points covers the space

Since $\mathcal{A}$ is a linear operator, for all $\widetilde{\mathbf{W}}$ with $\text{rank}(\widetilde{\mathbf{W}}) \leq 2r$,

$$(1-2\delta)\,\|\widetilde{\mathbf{W}}\|_F^2 \leq \|\mathcal{A}\big(\widetilde{\mathbf{W}}\big)\|_2^2 \leq (1+2\delta)\,\|\widetilde{\mathbf{W}}\|_F^2.$$

This means that $\mathcal{A}$ is $2r$-RIP with $\delta_{2r} = 2\delta$ when $m = O\left(\mu^2 d\bar{r}^2 r \log(36\sqrt{2}/\delta)/\delta^2\right)$. $\square$

Finally, we show that the sensing operator $\mathcal{T}$ is RIP on $(d_1+d_2)\times(d_1+d_2)$ PSD matrices of rank at most $2r$. Any of these PSD matrices can be written in form of $\mathcal{N} = \begin{pmatrix} \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T & \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T \\ \widetilde{\mathbf{V}}^T\widetilde{\mathbf{U}} & \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^T \end{pmatrix}$ where $\widetilde{\mathbf{U}} \in \mathbb{R}^{d_1\times 2r}$ and $\widetilde{\mathbf{V}} \in \mathbb{R}^{d_2\times 2r}$. We defined $\mathcal{T}$ such that $\mathcal{T}(\mathcal{N}) = 4\mathcal{A}(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T) + \big\|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T\big\|_F^2 + \big\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^T\big\|_F^2 - 2\big\|\widetilde{\mathbf{W}}\big\|_F^2$ where $\widetilde{\mathbf{W}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T$. Since

$$\|\mathcal{N}\|_F^2 = \big\|\mathbf{U}\mathbf{U}^T\big\|_F^2 + \big\|\mathbf{V}\mathbf{V}^T\big\|_F^2 + 2\big\|\widetilde{\mathbf{W}}\big\|_F^2,$$

if we have

$$(1-\delta)\,\big\|\widetilde{\mathbf{W}}\big\|_F^2 \leq \big\|\mathcal{A}(\widetilde{\mathbf{W}})\big\|_2^2 - \big\|\widetilde{\mathbf{W}}\big\|_F^2 \leq (1+\delta)\,\big\|\widetilde{\mathbf{W}}\big\|_F^2,$$

then it follows that

$$(1-2\delta)\,\big\|\mathcal{N}\big\|_F^2 \leq \big\|\mathcal{T}(\mathcal{N})\big\|_2^2 - \big\|\mathcal{N}\big\|_F^2 \leq (1+2\delta)\,\big\|\mathcal{N}\big\|_F^2.$$

Note that the deduction of the RIP of $\mathcal{T}$ from the RIP of $\mathcal{A}$ is thanks to the choice of the regularizer in (2.3).

2.4 Conclusion and Future Work

In this chapter, we discussed the geometric landscape of the inductive matrix completion (IMC) problem. The IMC model incorporates the side information in form of features of the row and column entities ($\mathbf{x}_i$'s and $\mathbf{y}_j$'s) and can be formulated as a low-rank matrix recovery problem where each observed entry of $\mathbf{M}^* = \mathbf{X}\mathbf{W}^*\mathbf{Y}$ is seen as a measurement of $\mathbf{W}^*$, that is $\mathbf{M}^*_{ij} = \mathbf{x}_i^T\mathbf{W}^*\mathbf{y}_j$. Motivated by the recovery guarantees of local search algorithms like AM for the factorized IMC problem [25], we study the optimization landscape of the factorized IMC problem. Using a framework developed by Ge et al. [5]

for matrix sensing problems, we show that, given $O(\max\{r^2, \log^2 n\}rd)$ observations, for the (regularized) factorized IMC problem *i*) there are no poor local minima, *ii*) the global minima satisfy $\mathbf{U}\mathbf{V}^T = \mathbf{W}^*$, *iii*) The Hessian at the saddle point has at least one negative eigenvalue.

This result shows that the recovery guarantees of AM in the IMC problem is not merely due to the algorithm and the geometry of the problem plays an important role. In fact, any algorithm, such as SGD, that can efficiently escape saddle points and find a local minimum can be used for solving the factorized IMC problem.

The IMC model has been studied extensively in the recent years. It has been employed in a variety of applications [48–51] and has been extended to more general settings such as IMC with noisy side information [27], high rank IMC [52], and non-linear IMC [53]. However, there are still many possible directions that have not yet been adequately explored. For example, tensor completion with side information is an area that although has received some attention in the recent years. A natural way to extend the IMC model to tensors is based on Tucker tensor decomposition. The Tucker decomposition factorizes an N-way tensor $\underline{\mathbf{M}} \in \mathbb{R}^{n_1, n_2, \cdots, d_N}$ in the following manner:

$$\underline{\mathbf{M}} = \underline{\mathbf{W}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \cdots \times_N \mathbf{U}_N,$$

where $\underline{\mathbf{W}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_N}$ denotes the core tensor, $\mathbf{U}_n \in \mathbb{R}^{d_n \times p_n}$ denote factor matrices along the n-th mode of $\underline{\mathbf{A}}$ for $n \in [N]$ and $\times_n$ denotes the mode-n product between a tensor and a matrix. Similar to inductive matrix completion, in many cases we have side information in form of feature matrices for the n mode entities, i.e. knowledge of the latent spaces of the underlying tensor. While there are a few recent works that explore this approach to incorporate side information into tensor completion [54–56], our understanding of this problem in terms of sample complexity, optimization landscape, and many other aspects is limited. While we leave extension the study of tensor completion with side information to future work, in the next chapter we employ some of the analytical tools that we use in Chapter 2 to provide recovery guarantees for a tensor problem, namely structured dictionary learning for tensor data.

Chapter 3

Learning Mixtures of Separable Dictionaries for Tensor Data

3.1 Introduction

Many data processing tasks such as feature extraction, data compression, classification, signal denoising, image inpainting, and audio source separation make use of data-driven sparse representations of data [8, 9, 13]. In many applications, these tasks are performed on data samples that are naturally structured as multiway arrays, also known as multidimensional arrays or tensors. Instances of *multidimensional* or *tensor* data include videos, hyperspectral images, tomographic images, and multiple-antenna wireless channels. Despite the ubiquity of tensor data in many applications, traditional data-driven sparse representation approaches disregard their multidimensional structure. This can result in sparsifying models with a large number of parameters. On the other hand, with the increasing availability of large data sets, it is crucial to keep sparsifying models reasonably small to ensure their scalable learning and efficient storage within devices such as smartphones and drones.

Our focus in this chapter is on learning of "compact" models that yield sparse representations of tensor data. To this end, we study *dictionary learning* (DL) for tensor data. The goal in DL, which is an effective and popular data-driven technique for obtaining sparse representations of data [8, 9, 13], is to learn a dictionary $\mathbf{D}$ such that every data sample can be approximated by a linear combination of a few atoms (columns) of $\mathbf{D}$. While DL has been widely studied, traditional DL approaches flatten tensor data and then employ methods designed for vector data [13, 57]. Such simplistic approaches disregard the multidimensional structure in tensor data and result in dictionaries with a large number of parameters. One intuitively expects, however,

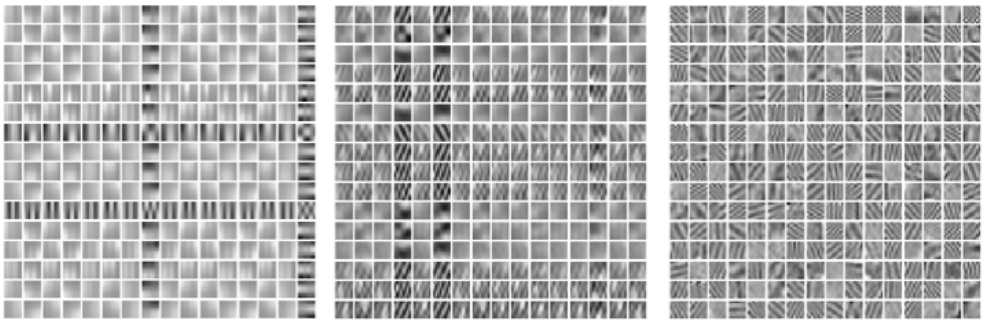


Figure 3.1: Dictionary atoms for representing RGB image `Barbara` for separation rank (left-to-right) 1, 4, and 256.

that dictionaries with smaller number of free parameters that exploit the correlation and structure along different tensor modes are likely to be more efficient with regards to storage requirements, computational complexity, and generalization performance, especially when training data are noisy or scarce.

To reduce the number of parameters in dictionaries for tensor data, and to better exploit the correlation among different tensor modes, some recent DL works have turned to tensor decompositions such as the Tucker decomposition [14] and CANDECOMP/PARAFAC decomposition (CPD) [15] for learning of "structured" dictionaries. The idea in *structured DL* for tensor data is to restrict the class of dictionaries during training to the one imposed by the tensor decomposition under consideration [58]. For example, structured DL based on the Tucker decomposition of N-way tensor data corresponds to the dictionary class in which any dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ consists of the Kronecker product [59] of N smaller *subdictionaries* $\{\mathbf{D}_n \in \mathbb{R}^{m_n \times p_n}\}_{n=1}^{N}$ [17, 60–64]. The resulting DL techniques in this instance are interchangeably referred to in the literature as *separable DL* or *Kronecker-structured DL* (KS-DL).

In terms of parameter counting, the advantages of KS-DL for tensor data are straightforward: the number of parameters needed to be estimated and stored for unstructured dictionary learning is $mp = \Pi_{n=1}^{N} m_n p_n$, whereas the KS-DL model requires only the sum of the subdictionary sizes $\sum_{n=1}^{N} m_n p_n$. Nonetheless, while existing KS-DL methods enjoy lower sample/computational complexity and better storage efficiency over unstructured DL [17], the KS-DL model makes a strong separability assumption among different modes of tensor data. Such an assumption can be overly restrictive for

many classes of data [18], resulting in an unfavorable tradeoff between model compactness and representation power.

Here, we overcome this limitation by proposing and studying a generalization of KS-DL that we interchangeably refer to as *learning a mixture of separable dictionaries* or *low separation rank DL* (LSR-DL). The separation rank of a matrix $\mathbf{A}$ is defined as the minimum number of KS matrices whose sum equals $\mathbf{A}$ [65, 66]. The LSR-DL model interpolates between the under-parameterized separable model (a special case of LSR-DL model with separation rank 1) and the over-parameterized unstructured model.[1] Figure 3.1 provides an illustrative example of the usefulness of LSR-DL, in which one learns a dictionary with a small separation rank: while KS-DL learns dictionary atoms that cannot reconstruct diagonal structures perfectly because of the abundance of horizontal/vertical (DCT-like) structures within them, LSR-DL also returns dictionary atoms with pronounced diagonal structures as the separation rank increases.[2]

3.1.1 Main Contributions

We first propose and analyze a generalization of the separable DL model—which we call a mixture of separable dictionaries model or LSR-DL model—that allows for better representation power than the separable model while having smaller number of parameters than standard DL. Our analysis assumes a generative model involving a true LSR dictionary for tensor data and investigates conditions under which the true dictionary is recoverable, up to a prescribed error, from training tensor data. Our first major set of LSR dictionary identifiability results are for the conventional optimization-based formulation of the DL problem [9], except that the search space is constrained to the class of dictionaries with maximum separation rank r (and individual mixture terms

[1] While KS-DL corresponds to Tucker decompostition, its generalization LSR-DL does not correspond to any of the well-known tensor factorizations.

[2] The results presented in this chapter have been published in Proceedings of 2017 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing [67], Proceedings of 2019 IEEE International Symposium on Information Theory [68], and IEEE Transactions on Signal Processing [69].

having bounded norms when $N \geq 3$ and $r \geq 2$).[3] Similar to conventional DL problems, this LSR-DL problem is nonconvex with multiple global minima. We therefore focus on *local identifiability* guarantees, meaning that a search algorithm initialized close enough to the true dictionary can recover that dictionary.[4] To this end, under certain assumptions on the generative model, we show that $\Omega\big(r(\sum_{n=1}^{N} m_n p_n)p^2\rho^{-2}\big)$ samples ensure existence of a local minimum of the constrained LSR-DL problem for Nth-order tensor data within a neighborhood of radius ρ around the true LSR dictionary.

Our initial local identifiability results are based on an analysis of a separation rank-constrained optimization problem that exploits a connection between LSR (resp., KS) matrices and low-rank (resp., rank-1) tensors. However, a result in tensor recovery literature [70] implies that finding the separation rank of a matrix is NP-hard. Our second main contribution is development and analysis of two different relaxations of the LSR-DL problem that are computationally tractable in the sense that they do not require explicit computation of the separation rank. The first formulation once again exploits the connection between LSR matrices and low-rank tensors and uses a convex regularizer to implicitly constrain the separation rank of the learned dictionary. The second formulation enforces the LSR structure on the dictionary by explicitly writing it as a summation of r KS matrices. Our analyses of the two relaxations once again involve conditions under which the true LSR dictionary is locally recoverable from training tensor data. We also provide extensive discussion in the sequel to compare and contrast the three sets of identifiability results for LSR dictionaries.

Our third main contribution is development of practical computational algorithms, which are based on the two relaxations of LSR-DL, for learning of an LSR dictionary in both batch and online settings. We then use these algorithms for learning of LSR dictionaries for both synthetic and real tensor data, which are afterward used in denoising and representation learning tasks. Numerical results obtained as part of these efforts help validate the usefulness of LSR-DL and highlight the different strengths and

[3]While we also provide identifiability results for LSR dictionaries without requiring the boundedness assumption, those results are only asymptotic in nature; see Section 3.3 for details.

[4]This is due to our choice of distance metric, which is the Frobenius norm.

weaknesses of the two LSR-DL relaxations and the corresponding algorithms.

3.1.2 Relation to Prior Work

Tensor decompositions [71, 72] have emerged as one of the main sets of tools that help avoid overparameterization of tensor data models in a variety of areas. These include deep learning, collaborative filtering, multilinear subspace learning, source separation, topic modeling, and many other works (see [73, 74] and references therein). But the use of tensor decompositions for reducing the (model and sample) complexity of dictionaries for tensor data has been addressed only recently.

There have been many works that provide theoretical analysis for the sample complexity of the conventional DL problem [75–78]. Among these, Gribonval et al. [77] focus on the local identifiability of the true dictionary underlying vectorized data using Frobenius norm as the distance metric. Shakeri et al. [17] extended this analysis for the sample complexity of the KS-DL problem for Nth-order tensor data. This analysis relies on expanding the objective function in terms of subdictionaries and exploiting the coordinate-wise Lipschitz continuity property of the objective function with respect to each subdictionary [17]. While this approach ensures the identifiability of the subdictionaries, it requires the dictionary coefficient vectors to follow the so-called *separable sparsity model* [79] and does not extend to the LSR-DL problem. In contrast, we provide local identifiability sample complexity results for the LSR-DL problem and two relaxations of it. Further, our identifiability results hold for coefficient vectors following the random sparsity model and the separable sparsity model.

In terms of computational algorithms, several works have proposed methods for learning KS dictionaries that rely on alternating minimization techniques to update the subdictionaries [61, 63, 79]. Among other works, Hawe et al. [60] employ a Riemannian conjugate gradient method combined with a nonmonotone line search for KS-DL. While they present the algorithm only for matrix data, its extension to higher-order tensor data is trivial. Schwab et al. [80] have also recently addressed the separable DL problem for matrix data; their contributions include a computational algorithm and global recovery guarantees. In terms of algorithms for LSR-DL, Dantas et al. [62]

proposed one of the first methods for matrix data that uses a convex regularizer to impose LSR on the dictionary. One of our batch algorithms, named STARK [81], also uses a convex regularizer for imposing LSR structure. In contrast to Dantas et al. [62], however, STARK can be used to learn a dictionary from tensor data of any order. The other batch algorithm we propose, named TeFDiL, learns subdictionaries of the LSR dictionary by exploiting the connection to tensor recovery and using tensor CPD. Recently, Dantas et al. [82] proposed an algorithm for learning an LSR dictionary for tensor data in which the dictionary update stage is a projected gradient descent algorithm that involves a CPD after every gradient step. In contrast, TeFDiL only requires a single CPD at the end of each dictionary update stage. Finally, while there exist a number of online algorithms for DL [57, 83, 84], the online algorithms developed in here are the first ones that enable learning of structured (either KS or LSR) dictionaries.

3.2 Preliminaries and Problem Statement

Notation and Definitions: We use underlined bold upper-case ($\underline{\mathbf{A}}$), bold upper-case ($\mathbf{A}$), bold lower-case ($\mathbf{a}$), and lower-case (a) letters to denote tensors, matrices, vectors, and scalars, respectively. For any integer p, we define $[p] \triangleq \{1, 2, \cdots, p\}$. We denote the j-th column of a matrix $\mathbf{A}$ by $\mathbf{a}_j$. For an $m \times p$ matrix $\mathbf{A}$ and an index set $\mathcal{J} \subseteq [p]$, we denote the matrix constructed from the columns of $\mathbf{A}$ indexed by $\mathcal{J}$ as $\mathbf{A}_{\mathcal{J}}$. We denote by $(\mathbf{A}_n)_{n=1}^N$ an N-tuple $(\mathbf{A}_1, \cdots, \mathbf{A}_N)$, while $\{\mathbf{A}_n\}_{n=1}^N$ represents the set $\{\mathbf{A}_1, \cdots, \mathbf{A}_N\}$. We drop the range indicators if they are clear from the context.

Norms and inner products: We denote by $\|\mathbf{v}\|_p$ the ℓ_p norm of vector $\mathbf{v}$, while we use $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_F$, and $\|\mathbf{A}\|_{\text{tr}}$ to denote the spectral, Frobenius, and trace (nuclear) norms of matrix $\mathbf{A}$, respectively. Moreover, $\|\mathbf{A}\|_{2,\infty} \triangleq \max_j \|\mathbf{a}_j\|_2$ is the *max column norm* and $\|\mathbf{A}\|_{1,1} \triangleq \sum_j \|\mathbf{a}_j\|_1$. We define the inner product of two tensors (or matrices) $\underline{\mathbf{A}}$ and $\underline{\mathbf{B}}$ as $\langle \underline{\mathbf{A}}, \underline{\mathbf{B}} \rangle \triangleq \langle \text{vec}(\underline{\mathbf{A}}), \text{vec}(\underline{\mathbf{B}}) \rangle$ where $\text{vec}(\cdot)$ is the vectorization operator. The Euclidean distance between two tuples of the same size is defined as $\left\|(\mathbf{A}_n)_{n=1}^N - (\mathbf{B}_n)_{n=1}^N\right\|_F \triangleq \sqrt{\sum_{n=1}^N ||\mathbf{A}_n - \mathbf{B}_n||_F^2}$.

Kronecker product: We denote by $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{m_1 m_2 \times p_1 p_2}$ the Kronecker product of

matrices $\mathbf{A} \in \mathbb{R}^{m_1 \times p_1}$ and $\mathbf{B} \in \mathbb{R}^{m_2 \times p_2}$. We use $\bigotimes_{n=1}^{N} \mathbf{A}_i \triangleq \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_N$ for the Kronecker product of N matrices. We drop the range indicators when there is no ambiguity. We call a matrix a (N-th order) Kronecker-structured (KS) matrix if it is a Kronecker product of $N \geq 2$ matrices.

Definitions for matrices: For a matrix $\mathbf{D}$ with unit ℓ_2-norm columns, we define the *cumulative coherence* μ_s as $\mu_s \triangleq \max_{|\mathcal{J}| \leq s} \max_{j \notin \mathcal{J}} \|\mathbf{D}_{\mathcal{J}}^T \mathbf{D}_j\|_1$. We say a matrix $\mathbf{D}$ satisfies the *s-restricted isometry property* (s-RIP) with constant δ_s if for any $\mathbf{v} \in \mathbb{R}^s$ and any $\mathcal{J} \subseteq [p]$ with $|\mathcal{J}| \leq s$, we have $(1-\delta_s)\|\mathbf{v}\|_2^2 \leq \|\mathbf{D}_{\mathcal{J}}\mathbf{v}\|_2^2 \leq (1+\delta_s)\|\mathbf{v}\|_2^2$.

Definitions for tensors: We briefly present required tensor definitions here: see Kolda and Bader [71] for more details. The mode-n unfolding matrix of $\underline{\mathbf{A}}$ is denoted by $\mathbf{A}_{(n)}$, where each column of $\mathbf{A}_{(n)}$ consists of the vector formed by fixing all indices of $\underline{\mathbf{A}}$ except the one in the nth-order. We denote the outer product (tensor product) of vectors by $\circ$, while $\times_n$ denotes the mode-n product between a tensor and a matrix. An N-way tensor is rank-1 if it can be written as outer product of N vectors: $\mathbf{v}_1 \circ \cdots \circ \mathbf{v}_N$. Throughout this chapter, by the rank of a tensor, rank($\underline{\mathbf{A}}$), we mean the CP-rank of $\underline{\mathbf{A}}$, the minimum number of rank-1 tensors that construct $\underline{\mathbf{A}}$ as their sum. The *CP decomposition* (CPD), decomposes a tensor into sum of its rank-1 tensor components. The *Tucker decomposition* factorizes an N-way tensor $\underline{\mathbf{A}} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_N}$ as $\underline{\mathbf{A}} = \underline{\mathbf{X}} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times_3 \cdots \times_N \mathbf{D}_N$, where $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_N}$ denotes the core tensor and $\mathbf{D}_n \in \mathbb{R}^{m_n \times p_n}$ denote factor matrices along the n-th mode of $\underline{\mathbf{A}}$ for $n \in [N]$.

Notations for functions and spaces: We denote the element-wise sign function by $\text{sgn}(\cdot)$. For any function $f(\mathbf{x})$, we define the difference $\Delta f(\mathbf{x}_1; \mathbf{x}_2) \triangleq f(\mathbf{x}_1) - f(\mathbf{x}_2)$. We denote by $\mathcal{U}_{m \times p}$ the Euclidean unit sphere: $\mathcal{U}_{m \times p} \triangleq \{\mathbf{D} \in \mathbb{R}^{m \times p} | \|\mathbf{D}\|_F = 1\}$. We also denote the Euclidean sphere with radius α by $\alpha\mathcal{U}_{m \times p}$. The oblique manifold in $\mathbb{R}^{m \times p}$ is the manifold of matrices with unit-norm columns: $\mathcal{D}_{m \times p} \triangleq \{\mathbf{D} \in \mathbb{R}^{m \times p} | \forall j \in [p],\ \mathbf{d}_j^T \mathbf{d}_j = 1\}$. We drop the dimension subscripts and use only $\mathcal{D}$ when there is no ambiguity. The covering number of a set $\mathcal{A}$ with respect to a norm $\|\cdot\|_*$, denoted by $\mathcal{N}_*(\mathcal{A}, \epsilon)$, is the minimum number of balls of $*$-norm radius ϵ needed to cover $\mathcal{A}$.

Dictionary Learning Setup: In dictionary learning (DL) for vector data, we

assume observations $\mathbf{y} \in \mathbb{R}^m$ are generated according to the following model:

$$\mathbf{y} = \mathbf{D}^0 \mathbf{x}^0 + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\mathbf{D}^0 \in \mathcal{D}_{m\times p} \subset \mathbb{R}^{m\times p}$ is the true underlying dictionary, $\mathbf{x}^0 \in \mathbb{R}^p$ is a randomly generated sparse coefficient vector, and $\boldsymbol{\epsilon} \in \mathbb{R}^m$ is the underlying noise vector. The goal in DL is to recover the true dictionary given the noisy observations $\mathbf{Y} \triangleq \{\mathbf{y}_l\}_{l=1}^L$ that are independent realizations of (3.1). The ideal objective is to solve the statistical risk minimization problem

$$\min_{\mathbf{D}\in\mathcal{C}} \; f_{\mathcal{P}}(\mathbf{D}) \triangleq \mathbb{E}_{\mathbf{y}\sim\mathcal{P}} \; f_{\mathbf{y}}(\mathbf{D}), \tag{3.2}$$

where $\mathcal{P}$ is the underlying distribution of the observations, $\mathcal{C} \subseteq \mathcal{D}_{m\times p}$ is the dictionary class, typically selected for vector data to be the same as the oblique manifold, and

$$f_{\mathbf{y}}(\mathbf{D}) \triangleq \inf_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1. \tag{3.3}$$

However, since we have access to the distribution $\mathcal{P}$ only through noisy observations drawn from this distribution, we resort to solving the following empirical risk minimization problem as a proxy for Problem (3.2):

$$\min_{\mathbf{D}\in\mathcal{C}} \; F_{\mathbf{Y}}(\mathbf{D}) \triangleq \frac{1}{L} \sum\nolimits_{l=1}^{L} f_{\mathbf{y}_l}(\mathbf{D}). \tag{3.4}$$

Dictionary Learning for Tensor Data: To represent tensor data, conventional DL approaches vectorize tensor data samples and treat them as one-dimensional arrays. One way to explicitly account for the tensor structure in data is to use the Kronecker-structured DL (KS-DL) model, which is based on the Tucker decomposition of tensor data. In the KS-DL model, we assume that observations $\underline{\mathbf{Y}}_l \in \mathbb{R}^{m_1\times\cdots\times m_N}$ are generated according to

$$\underline{\mathbf{Y}}_l = \underline{\mathbf{X}}_l^0 \times_1 \mathbf{D}_1^0 \times_2 \mathbf{D}_2^0 \times_3 \cdots \times_N \mathbf{D}_N^0 + \underline{\mathcal{E}}_l, \tag{3.5}$$

where $\{\mathbf{D}_n^0 \in \mathbb{R}^{m_n \times p_n}\}_{n=1}^N$ are generating *subdictionaries*, and $\underline{\mathbf{X}}_l^0$ and $\underline{\mathcal{E}}_l$ are the coefficient and noise tensors, respectively. Equivalently, the generating model (3.5) can be stated for $\mathbf{y}_l \triangleq \text{vec}(\underline{\mathbf{Y}}_l)$ as:

$$\mathbf{y}_l = \left(\mathbf{D}_N^0 \otimes \mathbf{D}_{N-1}^0 \otimes \cdots \otimes \mathbf{D}_1^0\right) \mathbf{x}_l^0 + \boldsymbol{\epsilon}_l, \tag{3.6}$$

where $\mathbf{x}_l^0 \triangleq \text{vec}(\underline{\mathbf{X}}^0{}_l)$ and $\boldsymbol{\epsilon}_l \triangleq \text{vec}(\underline{\mathcal{E}}_l)$ [71]. This is the same as the unstructured model $\mathbf{y}_l = \mathbf{D}^0\mathbf{x}_l^0 + \boldsymbol{\epsilon}_l$ with the additional condition that the generating dictionary is a Kronecker product of N subdictionaries. As a result, in the KS-DL problem, the constraint set in (3.4) becomes $\mathcal{C} = \mathcal{K}_{\mathbf{m},\mathbf{p}}^N$, where $\mathcal{K}_{\mathbf{m},\mathbf{p}}^N \triangleq \{\mathbf{D} \in \mathcal{D}_{m\times p} | \mathbf{D} = \bigotimes_{n=1}^N \mathbf{D}_n,\ \mathbf{D}_n \in \mathbb{R}^{m_n \times p_n}\}$ is the set of KS matrices with unit-norm columns and $\mathbf{m}$ and $\mathbf{p}$ are vectors containing m_n's and p_n's, respectively.[5]

In summary, the structure in tensor data is exploited in the KS-DL model by assuming the dictionary is "separable" into subdictionaries for each mode. However, as discussed earlier, this separable model is rather restrictive. Instead, we generalize the KS-DL model using the notion of *separation rank*.[6]

Definition 4. *The separation rank* $\mathfrak{R}_{\mathbf{m},\mathbf{p}}^N(\cdot)$ *of a matrix* $\mathbf{A} \in \mathbb{R}^{\Pi_n m_n \times \Pi_n p_n}$ *is the minimum number* r *of* N*th-order KS matrices* $\mathbf{A}^k = \bigotimes_{n=1}^N \mathbf{A}_n^k$ *such that* $\mathbf{A} = \sum_{k=1}^r \bigotimes_{n=1}^N \mathbf{A}_n^k$, *where* $\mathbf{A}_n^k \in \mathbb{R}^{m_n \times p_n}$.

The KS-DL model corresponds to dictionaries with separation rank 1. We instead propose the *low separation rank (LSR)* DL model in which the separation rank of the underlying dictionary is relatively small so that $1 \leq \mathfrak{R}_{\mathbf{m},\mathbf{p}}(\mathbf{D}^0) \ll \min\{m, p\}$. This generalizes the KS-DL model to a generating dictionary of the form $\mathbf{D}^0 = \sum_{k=1}^r [\mathbf{D}_N^k]^0 \otimes [\mathbf{D}_{N-1}^k]^0 \otimes \cdots \otimes [\mathbf{D}_1^k]^0$, where r is the separation rank of $\mathbf{D}^0$. Consequently, defining $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r} \triangleq \{\mathbf{D} \in \mathcal{D}_{m\times p} | \mathfrak{R}_{\mathbf{m},\mathbf{p}}^N(\mathbf{D}) \leq r\}$, the empirical *rank-constrained LSR-DL problem* is

$$\min_{\mathbf{D} \in \mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}} F_{\mathbf{Y}}(\mathbf{D}). \tag{3.7}$$

[5] We have changed the indexing of subdictionaries for ease of notation.

[6] The term was introduced in [66] for $N = 2$ (see also [65]).

However, the analytical tools at our disposal require the constraint set in (3.7) to be closed, which we show does not hold for $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ when $N \geq 3$ and $r \geq 2$. In that case, we instead analyze (3.7) with $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ replaced by (i) closure of $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ and (ii) a certain closed subset of $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$. We refer the reader to Section 3.3 for further discussion.

In our study of the LSR-DL model, which includes the KS-DL model as a special case, we use a correspondence between KS matrices and rank-1 tensors, stated in Lemma 2 below, which allows us to leverage techniques and results in the tensor recovery literature to analyze the LSR-DL problem and develop tractable algorithms. (This correspondence was first exploited in our earlier work [81].)

Lemma 2. *Any Nth-order Kronecker-structured matrix $\mathbf{A} = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_N$ can be rearranged as a rank-1, Nth-order tensor $\underline{\mathbf{A}}^{\pi} = \mathbf{a}_N \circ \cdots \circ \mathbf{a}_2 \circ \mathbf{a}_1$ with $\mathbf{a}_n \triangleq \mathrm{vec}(\mathbf{A}_n)$.*

The proof of Lemma 2 (details of the rearrangement procedure) is provided in the Appendix (Section 6.1). It follows immediately from Lemma 2 that if $\mathbf{D} = \sum_{k=1}^{r} \mathbf{D}_1^k \otimes \mathbf{D}_2^k \otimes \cdots \otimes \mathbf{D}_N^k$, then we can rearrange matrix $\mathbf{D}$ into the tensor $\underline{\mathbf{D}}^{\pi} = \sum_{k=1}^{r} \mathbf{d}_N^k \circ \mathbf{d}_{N-1}^k \circ \cdots \circ \mathbf{d}_1^k$, where $\mathbf{d}_n^k = \mathrm{vec}(\mathbf{D}_n^k)$. Therefore, we have the following equivalence:

$$\mathfrak{R}_{\mathbf{m},\mathbf{p}}^{N}(\mathbf{D}) \leq r \iff \mathrm{rank}(\underline{\mathbf{D}}^{\pi}) \leq r.$$

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 [70] and thus so is finding the separation rank of a matrix. This makes Problem (3.7) in its current form (and its variants) intractable. To overcome this, we introduce two tractable relaxations to the rank-constrained Problem (3.7) 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

$$\min_{\mathbf{D} \in \mathcal{D}_{m \times p}} F_{\mathbf{Y}}^{\mathrm{reg}}(\mathbf{D}) \tag{3.8}$$

with $F_{\mathbf{Y}}^{\mathrm{reg}}(\mathbf{D}) \triangleq \frac{1}{L} \sum_{l=1}^{L} f_{\mathbf{y}_l}(\mathbf{D}) + \lambda_1 g_1(\underline{\mathbf{D}}^{\pi})$, where $f_{\mathbf{y}}(\mathbf{D})$ is described in (3.3) and

$g_1(\underline{\mathbf{D}}^\pi)$ 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 subdictionaries. The resulting empirical risk minimization problem is

$$\min_{\{\mathbf{D}_n^k\}:\ \sum_{k=1}^r \bigotimes_{n=1}^N \mathbf{D}_n^k \in \mathcal{D}_{m\times p}} F_{\mathbf{Y}}^{\text{fac}}\big(\{\mathbf{D}_n^k\}\big), \tag{3.9}$$

where $F_{\mathbf{Y}}^{\text{fac}}(\{\mathbf{D}_n^k\}) \triangleq \frac{1}{L}\sum_{l=1}^L f_{\mathbf{y}_l}^{\text{fac}}(\{\mathbf{D}_n^k\})$ with

$$f_{\mathbf{y}}^{\text{fac}}(\{\mathbf{D}_n^k\}) \triangleq \inf_{\mathbf{x}\in\mathbb{R}^p} \Big\|\mathbf{y} - \big(\sum\nolimits_{k=1}^r \bigotimes\nolimits_{n=1}^N \mathbf{D}_n^k\big)\mathbf{x}\Big\|^2 + \lambda\,\|\mathbf{x}\|_1\,,$$

and the terms $\bigotimes_{n=1}^N \mathbf{D}_n^k$ are constrained as $\|\bigotimes_{n=1}^N \mathbf{D}_n^k\|_F \leq c$ for some positive constant c when $N \geq 3$ and $r \geq 2$.

In the rest of this chapter, we study the problem of identifying the true underlying LSR-DL dictionary by analyzing the LSR-DL Problems (3.7)–(3.9) introduced in this section and developing algorithms to solve Problems (3.8) and (3.9) in both batch and online settings. Note that while Problem (3.7) (and its variants when $N \geq 3$ and $r \geq 2$) cannot be explicitly solved because of its NP-hardness, identifiability analysis of this problem—provided in Section 3.3—provides the basis for the analysis of tractable Problems (3.8) and (3.9), provided in Section 3.4. To improve the readability of our notation-heavy discussions and analysis, we have provided a table of notations (Table 3.1) for easy access to definitions of the most commonly used notation.

3.3 Identifiability in the Rank-constrained LSR-DL Problem

In this section, we derive conditions under which a dictionary $\mathbf{D}^0 \in \mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ is identifiable as a solution to either the separation rank-constrained problem in (3.7) or a slight variant of (3.7) when $N \geq 3$ and $r \geq 2$. Specifically, we show that under certain assumptions on the generative model, there is at least one local minimum $\mathbf{D}^*$ of either Problem (3.7) or one of its variants that is "close" to the underlying dictionary $\mathbf{D}^0$. Notwithstanding the fact that no efficient algorithm exists to solve the intractable

Table 3.1: Table of commonly used notation in Chapter 3

Notation	Definition
m, p	$\prod_{n=1}^{N} m_n$, $\prod_{n=1}^{N} p_n$
$\mathbf{m}, \mathbf{p}$	$(m_n)_{n=1}^{N}$, $(p_n)_{n=1}^{N}$
$\mathcal{N}_*(\mathcal{A}, \epsilon)$	Covering number of set $\mathcal{A}$ w.r.t. norm $*$
$\mathfrak{R}^N_{\mathbf{m},\mathbf{p}}(\mathbf{D})$	Separation rank of matrix $\mathbf{D}$
$\mathcal{D}_{m\times p}$	Oblique manifold in $\mathbb{R}^{m\times p}$
$\mathcal{U}_{m\times p}$	Euclidean unit sphere in $\mathbb{R}^{m\times p}$
$\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$	Set of LSR matrices: $\{\mathbf{D} \in \mathbb{R}^{m\times p} \vert \mathfrak{R}^N_{\mathbf{m},\mathbf{p}}(\mathbf{D}) \leq r\}$
$\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$	$\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}} \cap \mathcal{D}_{m\times p}$
$\mathcal{K}^{N}_{\mathbf{m},\mathbf{p}}$	$\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ with $r = 1$: Set of KS matrices on $\mathcal{D}_{m\times p}$
$\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}$	$\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ with $N = 2$
${}^{c}\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$	$\{\mathbf{D} \in \mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}} \vert \vert\vert \bigotimes \mathbf{D}^k_n \vert\vert_F \leq c, c > 0\}$
$\overline{\mathcal{K}}^{N,r}_{\mathbf{m},\mathbf{p}}$	Closure of $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$
$\mathcal{C}$	Compact constraint set in LSR-DL problem: one of $\mathcal{K}^{N}_{\mathbf{m},\mathbf{p}}$, $\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}$, ${}^{c}\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$, or $\overline{\mathcal{K}}^{N,r}_{\mathbf{m},\mathbf{p}}$
$\mathcal{B}_\rho$	$\{\mathbf{D} \in \mathcal{C} \vert \vert\mathbf{D} - \mathbf{D}^0\vert_F \leq \rho\}$
$\Delta f(\mathbf{x}_1; \mathbf{x}_2)$	$f(\mathbf{x}_1) - f(\mathbf{x}_2)$
$f_{\mathbf{y}}(\mathbf{D})$	$\inf_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{2} \left\|\mathbf{y} - \mathbf{D}\mathbf{x}\right\|_2^2 + \lambda \|\mathbf{x}\|_1$
$f_{\mathcal{P}}(\mathbf{D})$	$\mathbb{E}_{\mathbf{y}\sim\mathcal{P}}\ f_{\mathbf{y}}(\mathbf{D})$
$\Delta f_{\mathcal{P}}(\rho)$	$\inf_{\mathbf{D}\in\partial\mathcal{B}_\rho} \Delta f_{\mathcal{P}}(\mathbf{D}; \mathbf{D}^0)$
$F_{\mathbf{Y}}(\mathbf{D})$	$\frac{1}{L}\sum_{l=1}^{L} f_{\mathbf{y}_l}(\mathbf{D})$
$F^{\text{reg}}_{\mathbf{Y}}(\mathbf{D})$	$\frac{1}{L}\sum_{l=1}^{L} f_{\mathbf{y}_l}(\mathbf{D}) + \lambda_1 g_1(\underline{\mathbf{D}}^\pi)$
$f^{\text{fac}}_{\mathbf{y}}(\{\mathbf{D}^k_n\})$	$\inf_{\mathbf{x}\in\mathbb{R}^p} \left\|\mathbf{y} - \left(\sum_{k=1}^{r} \bigotimes_{n=1}^{N} \mathbf{D}^k_n\right)\mathbf{x}\right\|_2^2 + \lambda \left\|\mathbf{x}\right\|_1$
$F^{\text{fac}}_{\mathbf{Y}}(\{\mathbf{D}^k_n\})$	$\frac{1}{L}\sum_{l=1}^{L} f^{\text{fac}}_{\mathbf{y}_l}(\{\mathbf{D}^k_n\})$

Problem (3.7), this identifiability result is important in that it lays the foundation for the local identifiability results in tractable Problems (3.8) and (3.9).

Generative Model: Let $\mathbf{D}^0 \in \mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ be the underlying dictionary. Each tensor data sample $\underline{\mathbf{Y}} \in \mathbb{R}^{m_1\times m_2\times\cdots\times m_N}$ in its vectorized form is *independently* generated using a linear combination of $s \ll p$ atoms of dictionary $\mathbf{D}^0$ with added noise: $\mathbf{y} \triangleq \text{vec}(\underline{\mathbf{Y}}) = \mathbf{D}^0\mathbf{x}^0 + \boldsymbol{\epsilon}$, where $\left\|\mathbf{x}^0\right\|_0 \leq s$. Specifically, s atoms of $\mathbf{D}^0$ are selected uniformly at random, defining the support $\mathcal{J} \subset [p]$. Then, we draw a random sparse coefficient vector $\mathbf{x}^0 \in \mathbb{R}^p$ supported on $\mathcal{J}$. We state further assumptions on our model similar to the prior works [17, 77].

Assumption 1 (Coefficient Distribution)**.** *We assume that*

i) the nonzero elements of $\mathbf{x}^0$ *are zero-mean and uncorrelated:*

$$\mathbb{E}\{\mathbf{x}^0_{\mathcal{J}}[\mathbf{x}^0_{\mathcal{J}}]^T|\mathcal{J}\} = \mathbb{E}\{x^2\} \cdot \mathbf{I}_s,$$

ii) *the nonzero elements of* $\mathbf{s}^0 \triangleq \operatorname{sgn}(\mathbf{x}^0)$ *are zero-mean and uncorrelated:* $\mathbb{E}\{\mathbf{s}^0_{\mathcal{J}}[\mathbf{s}^0_{\mathcal{J}}]^T|\mathcal{J}\} = \mathbf{I}_s$,

iii) $\mathbf{x}^0$ *and* $\mathbf{s}^0$ *are uncorrelated:* $\mathbb{E}\{\mathbf{s}^0_{\mathcal{J}}[\mathbf{x}^0_{\mathcal{J}}]^T|\mathcal{J}\} = \mathbb{E}\{|x|\} \cdot \mathbf{I}_s$,

iv) $\mathbf{x}^0$ *has bounded norm almost surely:* $\left\|\mathbf{x}^0\right\|_2 \leq M_x$ *with probability* 1,

v) *nonzero elements of* $\mathbf{x}^0$ *are far from zero almost surely:* $\min_{j\in\mathcal{J}} |\mathbf{x}^0_j| \geq \underline{x}$ *with probability* 1.

Assumption 2 (Noise Distribution)**.** *We make the following assumptions on the distribution of the noise* $\boldsymbol{\epsilon}$*:*

i) *the elements of* $\boldsymbol{\epsilon}$ *are zero-mean and uncorrelated:* $\mathbb{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\mathcal{J}\} = \mathbb{E}\{\epsilon^2\} \cdot \mathbf{I}$,

ii) $\boldsymbol{\epsilon}$ *is uncorrelated with* $\mathbf{x}^0$ *and* $\mathbf{s}^0$*:* $\mathbb{E}\{\mathbf{x}^0\boldsymbol{\epsilon}^T|\mathcal{J}\} = \mathbb{E}\{\mathbf{s}^0\boldsymbol{\epsilon}^T|\mathcal{J}\} = 0$,

iii) $\boldsymbol{\epsilon}$ *has bounded norm almost surely:* $\|\boldsymbol{\epsilon}\|_2 \leq M_\epsilon$ *with probability* 1.

Note that Assumptions 1-iv and 2-iii imply the magnitude of $\mathbf{y}$ is bounded: $\|\mathbf{y}\|_2 \leq M_y$. Next, we define positive parameters $\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\{|x|\}}$, $C_{\max} \triangleq \frac{2\mathbb{E}|x|}{7M_x}\left(1 - 2\mu_s(\mathbf{D}^0)\right)$,

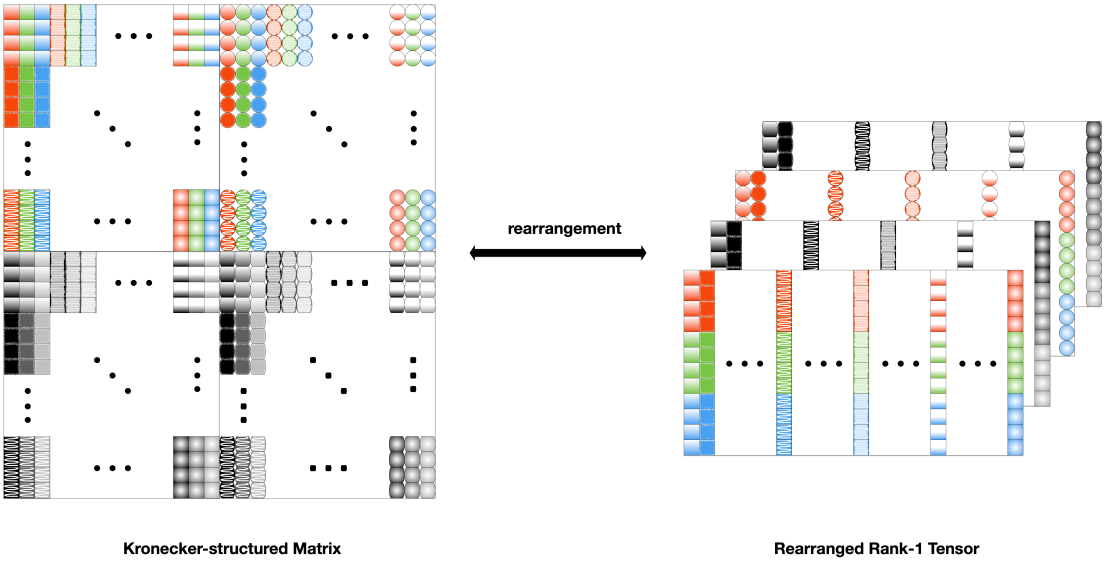


Figure 3.2: Example of rearranging a Kronecker structured matrix ($N = 3$) into a third order rank-1 tensor.

and $C_{\min} \triangleq 24\frac{\mathbb{E}\{|x|\}^2}{\mathbb{E}\{x^2\}}\left(\left\|\mathbf{D}^0\right\|_2+1\right)^2\frac{s}{p}\left\|[\mathbf{D}^0]^T\mathbf{D}^0-\mathbf{I}\right\|_F$ for ease of notation. We use the following assumption, similar to Gribonval et al. [77, Thm. 1].

Assumption 3. *Assume* $C_{\min} \leq C_{\max}$, $\lambda \leq \underline{x}/4$, $s \leq \frac{p}{16\left(\|\mathbf{D}^0\|_2+1\right)^2}$, $\mu_s(\mathbf{D}^0) \leq 1/4$, *and the noise is relatively small in the sense that* $\frac{M_\epsilon}{M_x} < \frac{7}{2}\left(C_{\max}-C_{\min}\right)\bar{\lambda}$.

Our Approach: In our analysis of the separation rank-constrained LSR-DL problem, we will alternate between four different constraint sets that are related to our dictionary class $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$, namely, $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{2,r}$, $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N}$, the closure $\overline{\mathcal{K}}_{\mathbf{m},\mathbf{p}}^{N,r} \triangleq \text{cl}(\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r})$ of $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ under the Frobenius norm, and a closed subset of $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$, defined as ${}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r} \triangleq \{\mathbf{D} \in \mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r} | \|\bigotimes \mathbf{D}_n^k\|_F \leq c, c>0\}$. We often use the generic notation $\mathcal{C}$ for the constraint set when our discussion is applicable to more than one of these sets.

We want to find conditions that imply the existence of a local minimum of the problem $\min_{\mathbf{D}\in\mathcal{C}} F_{\mathbf{Y}}(\mathbf{D})$ within a ball of radius ρ around the true dictionary $\mathbf{D}^0 \in \mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$:

$$\mathcal{B}_\rho \triangleq \{\mathbf{D} \in \mathcal{C} | \left\|\mathbf{D}-\mathbf{D}^0\right\|_F \leq \rho\} \tag{3.10}$$

for some small $\rho > 0$. To this end, we first show that the expected risk function $f_{\mathcal{P}}(\mathbf{D})$ in (3.2) has a local minimum in $\mathcal{B}_\rho$ for the LSR-DL constraint set $\mathcal{C}$.

To show that a local minimum of $f_{\mathcal{P}} : \mathcal{C} \mapsto \mathbb{R}$ exists in $\mathcal{B}_\rho$, we need to show that $f_{\mathcal{P}}(\mathbf{D})$ attains its minimum over $\mathcal{B}_\rho$ in the interior of $\mathcal{B}_\rho$.[7] We show this in two stages. First, we use the Weierstrass Extreme Value Theorem [85], which dictates that the continuous function $f_{\mathcal{P}}(\mathbf{D})$ attains a minimum in (or on the boundary of) $\mathcal{B}_\rho$ as long as $\mathcal{B}_\rho$ is a compact set. Therefore, we first investigate compactness of $\mathcal{B}_\rho$ in Section 3.3.1. Second, in order to be certain that the minimum of $f_{\mathcal{P}}(\mathbf{D})$ over $\mathcal{B}_\rho$ is a local minimum of $\mathbf{D} \in \mathcal{C} \mapsto f_{\mathcal{P}}(\mathbf{D})$, we show that $f_{\mathcal{P}}(\mathbf{D})$ cannot obtain its minimum over $\mathcal{B}_\rho$ on the boundary of $\mathcal{B}_\rho$, denoted by $\partial\mathcal{B}_\rho$. To this end, in Section 3.3.2 we derive conditions

[7] Having a minimum $\mathbf{D}^*$ on the boundary is not sufficient since the function might achieve lower values in the neighborhood of $\mathbf{D}^*$ outside $\mathcal{B}_\rho$.

that if $\partial\mathcal{B}_\rho$ is nonempty then[8]

$$\Delta f_{\mathcal{P}}(\rho) \triangleq \inf_{\mathbf{D}\in\partial\mathcal{B}_\rho} \Delta f_{\mathcal{P}}(\mathbf{D};\mathbf{D}^0) > 0, \tag{3.11}$$

which implies $f_{\mathcal{P}}(\mathbf{D})$ cannot achieve its minimum on $\partial\mathcal{B}_\rho$.

Finally, in Section 3.3.3 we use concentration of measure inequalities to relate $F_{\mathbf{Y}}(\mathbf{D})$ in (3.4) to $f_{\mathcal{P}}(\mathbf{D})$ and find the number of samples needed to guarantee (with high probability) that $F_{\mathbf{Y}}(\mathbf{D})$ also has a local minimum in the interior of $\mathcal{B}_\rho$.

3.3.1 Compactness of the Constraint Sets

When the constraint set $\mathcal{C}$ is a compact subset of the Euclidean space $\mathbb{R}^{m\times p}$, the subset $\mathcal{B}_\rho$ is also compact. Thus, we first investigate the compactness of the constraint set $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$. Since $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ is a bounded set, according to the Heine-Borel Theorem [85], it is a compact subset of $\mathbb{R}^{m\times p}$ if and only if it is closed. Also, $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ can be written as the intersection of $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}} \triangleq \{\mathbf{D}\in\mathbb{R}^{m\times p} | \mathfrak{R}^N_{\mathbf{m},\mathbf{p}}(\mathbf{D}) \leq r\}$ and the oblique manifold $\mathcal{D}$. In order for $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}} = \mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}} \cap \mathcal{D}$ to be closed, it suffices to show that $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ and $\mathcal{D}$ are closed. It is trivial to show $\mathcal{D}$ is closed; hence, we focus on whether $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ is closed.

In the following, we use the facts that the constraint $\mathfrak{R}^N_{\mathbf{m},\mathbf{p}}(\mathbf{D}) \leq r$ is equivalent to $\text{rank}(\underline{\mathbf{D}}^\pi) \leq r$ and that the rearrangement mapping that sends $\mathbf{D}$ to $\underline{\mathbf{D}}^\pi$ preserves topological properties of sets such as the distances between the set elements under the Frobenius norm. These facts allow us to translate the topological properties of tensor sets into properties of the structured matrices that we study here.

Lemma 3. *Let $N \geq 3$ and $r \geq 2$. Then, the set $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ is not closed. However, the set of KS matrices $\mathcal{L}^{N,1}_{\mathbf{m},\mathbf{p}}$ and the set $\mathcal{L}^{2,r}_{\mathbf{m},\mathbf{p}}$ are closed.*

Proof. Proposition 4.1 in De Silva and Lim [86] shows that the space of tensors of order $N \geq 3$ and rank $r \geq 2$ is not closed. The fact that the rearrangement process preserves topological properties of sets means that the same result holds for the set $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ with $N \geq 3$ and rank $r \geq 2$.

[8] If the boundary is empty, it is trivial that the infimum is attained in the interior of the set.

The proof for closedness of $\mathcal{L}^{N,1}_{\mathbf{m},\mathbf{p}}$ and $\mathcal{L}^{2,r}_{\mathbf{m},\mathbf{p}}$ follows from Propositions 4.2 and 4.3 in De Silva and Lim [86], which can be adopted here due to the relation between the sets of low-rank tensors and LSR matrices. □

To illustrate the non-closedness of $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ for $N \geq 3$ and $r \geq 2$ and motivate the use of the sets $\overline{\mathcal{K}}^{N,r}_{\mathbf{m},\mathbf{p}}$ and ${}^{c}\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ in lieu of $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$, we provide an example. Consider the sequence $\mathbf{D}_t := t\left(\mathbf{A}_1 + \frac{1}{t}\mathbf{B}_1\right) \otimes \left(\mathbf{A}_2 + \frac{1}{t}\mathbf{B}_2\right) \otimes \left(\mathbf{A}_3 + \frac{1}{t}\mathbf{B}_3\right)) - t\mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \mathbf{A}_3$ where $\mathbf{A}_i, \mathbf{B}_i \in \mathbb{R}^{m_i \times p_i}$ are linearly independent pairs. It is clear that $\mathfrak{R}^3_{\mathbf{m},\mathbf{p}}(\mathbf{D}_t) \leq 2$ for any t. The limit point of this sequence, however, is $\lim_{t\to\infty} \mathbf{D}_t = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \mathbf{B}_3 + \mathbf{A}_1 \otimes \mathbf{B}_2 \otimes \mathbf{A}_3 + \mathbf{B}_1 \otimes \mathbf{A}_2 \otimes \mathbf{B}_3$, which is a separation-rank-3 matrix. Therefore, the set $\mathcal{L}^{3,2}_{\mathbf{m},\mathbf{p}}$ is not closed.

The non-closedness of $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ means there exist sequences in $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ whose limit points are not in the set. Two possible solutions to circumvent this issue include: (i) use the closure of $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ as the constraint set, and (ii) eliminate such sequences from $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$. We discuss each solution in detail below.

Adding the limit points

We denote the closure of $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ by $\overline{\mathcal{L}}^{N,r}_{\mathbf{m},\mathbf{p}} \triangleq \mathrm{cl}(\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}})$. By slightly relaxing the constraint set in (3.7) to $\overline{\mathcal{L}}^{N,r}_{\mathbf{m},\mathbf{p}} \cap \mathcal{D}$, we can instead solve the following:

$$\min_{\mathbf{D} \in \overline{\mathcal{K}}^{N,r}_{\mathbf{m},\mathbf{p}}} F_{\mathbf{Y}}(\mathbf{D}), \tag{3.12}$$

where $\overline{\mathcal{K}}^{N,r}_{\mathbf{m},\mathbf{p}} = \overline{\mathcal{L}}^{N,r}_{\mathbf{m},\mathbf{p}} \cap \mathcal{D}$. Note that (i) a solution to (3.7) is a solution to (3.12) and (ii) a solution to (3.12) is either a solution to (3.7) or is arbitrarily close to a member of $\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$.[9]

[9]The first argument holds since if $F_{\mathbf{Y}}(\mathbf{D}^*) \leq F_{\mathbf{Y}}(\mathbf{D})$ for all $\mathbf{D} \in \mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$, by continuity it also holds for all $\mathbf{D} \in \overline{\mathcal{K}}^{N,r}_{\mathbf{m},\mathbf{p}}$. The second argument is trivial.

Eliminating the problematic sequences

In order to exclude the sequences $\mathbf{D}_t \to \mathbf{D}$ such that $\mathbf{D}_t \in \mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ for all t and $\mathbf{D} \notin \mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$, we first need to characterize them.

Lemma 4. *Assume* $\mathbf{D}_t \to \mathbf{D}$ *where* $\mathfrak{R}^N_{\mathbf{m},\mathbf{p}}(\mathbf{D}_t) \leq r$ *and* $\mathfrak{R}^N_{\mathbf{m},\mathbf{p}}(\mathbf{D}) > r$. *We can write* $\mathbf{D}_t = \sum_{k=1}^r \lambda_t^k \bigotimes_{n=1}^N [\mathbf{D}_n^k]_t$ *where* $\left\|[\mathbf{D}_n^k]_t\right\|_F = 1$. *Then,* $\max_k |\lambda_t^k| \to \infty$ *as* $t \to \infty$. *In fact, at least two of the coefficient sequences* λ_t^k *are unbounded.*

Proof. The rearrangement process allows us to borrow the results in Proposition 4.8 in De Silva and Lim [86] for tensors and apply them to LSR matrices. □

The following corollary of Lemma 4 suggests that one can exclude the problematic sequences from $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ by bounding the norm of individual KS (separation-rank-1) terms.

Corollary 1. *Consider the set* $\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}}$ *whose members can be written as* $\mathbf{D} = \sum_{k=1}^{r} \bigotimes_{n=1}^{N} \mathbf{D}_n^k$ *such that* $\mathbf{D}_n^k \in \mathbb{R}^{m_n \times p_n}$. *Then, for any* $c > 0$ *the set* ${}^c\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}} = \left\{\mathbf{D} \in \mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}} \middle| \left\|\bigotimes \mathbf{D}_n^k\right\|_F \leq c\right\}$ *is closed.*

We have now shown that the sets $\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}$, $\mathcal{K}^{N}_{\mathbf{m},\mathbf{p}} \triangleq \mathcal{K}^{N,1}_{\mathbf{m},\mathbf{p}}$, ${}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}} = {}^c\mathcal{L}^{N,r}_{\mathbf{m},\mathbf{p}} \cap \mathcal{D}$, and $\overline{\mathcal{K}}^{N,r}_{\mathbf{m},\mathbf{p}} = \overline{\mathcal{L}}^{N,r}_{\mathbf{m},\mathbf{p}} \cap \mathcal{D}$ are compact subsets of $\mathbb{R}^{m\times p}$. Next, we provide asymptotic identifiability results for these compact constraint sets.

3.3.2 Asymptotic Analysis for Dictionary Identifiability

Now that we have discussed the compactness of the relevant constraint sets, we are ready to show that the minimum of $f_{\mathbf{y}}(\mathbf{D})$ over $\mathcal{B}_\rho$, defined in (3.10), is not attained on $\partial\mathcal{B}_\rho$. This will complete our proof of existence of a local minimum of $f_{\mathcal{P}}(\mathbf{D})$ in $\mathcal{B}_\rho$. In our proof, we make use of a result in Gribonval et al. [77], presented here in Lemma 5.

Lemma 5 (Theorem 1 in Gribonval et al. [77])**.** *Consider the statistical DL Problem* (3.2) *with constraint set* $\mathcal{D}$. *Suppose the generating dictionary* $\mathbf{D}^0 \in \mathcal{D}$ *and Assumptions 1–3 hold. Then, for any* ρ *such that* $\bar{\lambda}C_{\min} < \rho \leq \bar{\lambda}C_{\max}$ *and* $\frac{M_\epsilon}{M_x} < \frac{7}{2}(\bar{\lambda}C_{max} - \rho)$, *we*

have

$$\Delta f_{\mathcal{P}}(\rho) \geq \frac{\mathbb{E}\{x^2\}}{8} \cdot \frac{s}{p} \cdot \rho \left(\rho - \bar{\lambda} C_{\min}\right) > 0. \tag{3.13}$$

for all $\mathbf{D} \in \mathcal{D}$ *such that* $\|\mathbf{D} - \mathbf{D}^0\|_F = \rho$.

Interested readers can find the detailed proof of Lemma 5 in Gribonval et al. [77]. The following theorem states our first identifiability result for the LSR-DL model.

Theorem 4. *Consider the statistical DL Problem* (3.2) *with constraint set* $\mathcal{C}$ *being either* $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{2,r}$, $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N}$, ${}^{c}\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ *or* $\overline{\mathcal{K}}_{\mathbf{m},\mathbf{p}}^{N,r}$. *Suppose the generating dictionary* $\mathbf{D}^0 \in \mathcal{C}$ *and Assumptions 1–3 hold. Then, for any* ρ *such that* $\bar{\lambda} C_{\min} < \rho < \bar{\lambda} C_{\max}$ *and* $\frac{M_\epsilon}{M_x} < \frac{7}{2}(\bar{\lambda} C_{max} - \rho)$, *the function* $\mathbf{D} \in \mathcal{C} \mapsto f_{\mathcal{P}}(\mathbf{D})$ *has a local minimum* $\mathbf{D}^*$ *such that* $\left\|\mathbf{D}^* - \mathbf{D}^0\right\|_F < \rho$.

Proof. Since $f_{\mathcal{P}}(\mathbf{D})$ is a continuous function and the ball $\mathcal{B}_\rho = \{\mathbf{D} \in \mathcal{C} | \left\|\mathbf{D} - \mathbf{D}^0\right\|_F \leq \rho\}$ is compact, by the extreme value theorem, $\mathbf{D} \in \mathcal{B}_\rho \mapsto f_{\mathcal{P}}(\mathbf{D})$ attains its infimum at a point in the ball. If this minimum is attained in the interior of $\mathcal{B}_\rho$ then it is a local minimum of $\mathbf{D} \in \mathcal{C} \mapsto f_{\mathcal{P}}(\mathbf{D})$. Therefore, a key ingredient of the proof is showing that $f_{\mathcal{P}}(\mathbf{D}) > f_{\mathcal{P}}(\mathbf{D}^0)$ for all $\mathbf{D} \in \partial\mathcal{B}_\rho$ if $\partial\mathcal{B}_\rho$ is nonempty. Lemma 5 states the conditions under which $f_{\mathcal{P}}(\mathbf{D}) > f_{\mathcal{P}}(\mathbf{D}^0)$ on $\partial\mathcal{S}_\rho$, where $\mathcal{S}_\rho \triangleq \left\{\mathbf{D} \in \mathcal{D} \;\middle|\; \left\|\mathbf{D} - \mathbf{D}^0\right\|_F \leq \rho\right\}$.

Since $\partial\mathcal{B}_\rho \subset \partial\mathcal{S}_\rho$, the result of Lemma 5 can be used for our problem as well, i.e. for any $\mathbf{D} \in \partial\mathcal{B}_\rho$, we have $f_{\mathcal{P}}(\mathbf{D}) > f_{\mathcal{P}}(\mathbf{D}^0)$, when $C_{\min}\bar{\lambda} < \rho < C_{\max}\bar{\lambda}$. It follows from this result together with the existence of the infimum of $f_{\mathcal{P}}(\mathbf{D}) : \mathcal{B}_\rho \mapsto \mathbb{R}$ in $\mathcal{B}_\rho$ that Problem (3.2) has a local minimum within a ball of radius ρ around the true dictionary $\mathbf{D}^0$. □

Next, we discuss finite sample identifiability of the true dictionary $\mathbf{D}^0$ for three of the constraint sets.

3.3.3 Sample Complexity for Dictionary Identifiability

We now derive the number of samples required to guarantee, with high probability, that $F_{\mathbf{Y}} : \mathcal{C} \mapsto \mathbb{R}$ has a local minimum at a point "close" to $\mathbf{D}^0$ when the constraint set

$\mathcal{C}$ is either $\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}$, $\mathcal{K}^{N}_{\mathbf{m},\mathbf{p}}$, or ${}^{c}\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ for $N \geq 3$ and $r \geq 2$. First, we use concentration of measure inequalities based on the covering number of the dictionary class $\mathcal{C} \subset \mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ to show that the empirical loss $F_{\mathbf{Y}}(\mathbf{D})$ uniformly converges to its expectation $f_{\mathcal{P}}(\mathbf{D})$ with high probability. This is formalized below.

Lemma 6 (Theorem 1 and Lemma 11, Gribonval et al. [87])**.** *Consider the empirical DL Problem* (3.4) *and suppose Assumptions 1 and 2 are satisfied. For any $u \geq 0$ and constants $c_1 \geq M_y^2/\sqrt{8}$ and $c_2 \geq \max(1, \log c_0\sqrt{8}M_y)$, with probability at least $1-2e^{-u}$ we have*

$$\sup_{\mathbf{D}\in\mathcal{C}} |F_{\mathbf{Y}}(\mathbf{D}) - f_p(\mathbf{D})| \leq 3c_1\sqrt{\frac{c_2\nu\log L}{L}} + c_1\sqrt{\frac{c_2\nu+u}{L}}, \tag{3.14}$$

where ν is such that $\mathcal{N}_{2,\infty}(\mathcal{C},\epsilon) = \left(\frac{c_0}{\epsilon}\right)^{\nu}$.

Define $\eta_L \triangleq 3c_1\sqrt{\frac{c_2\nu\log L}{L}} + c_1\sqrt{\frac{c_2\nu+u}{L}}$. It follows from (3.14) that with high probability (w.h.p.),

$$\Delta F_{\mathbf{Y}}(\mathbf{D};\mathbf{D}^0) \geq \Delta f_{\mathcal{P}}(\mathbf{D};\mathbf{D}^0) - 2\eta_L, \tag{3.15}$$

for all $\mathbf{D} \in \mathcal{C}$. Therefore, when $\eta_L < \Delta f_{\mathcal{P}}(\mathbf{D};\mathbf{D}^0)/2$ for all $\mathbf{D} \in \partial\mathcal{B}_\rho$, we have $\Delta F_{\mathbf{Y}}(\mathbf{D};\mathbf{D}^0) > 0$ for all $\mathbf{D} \in \partial\mathcal{B}_\rho$. In this case, we can use similar arguments as in the asymptotic analysis to show that $F_{\mathbf{Y}} : \mathcal{C} \to \mathbb{R}$ has a local minimum at a point in the interior of $\mathcal{B}_\rho$. Hence, our focus in this section is on finding the sample complexity L required to guarantee that $\eta_L \leq \Delta f_{\mathcal{P}}(\rho)/2$ w.h.p. We begin with characterization of covering numbers of the three constraint sets, which may also be of independent interest to some readers.

Covering Numbers: The covering number of the set $\mathcal{K}^{N}_{\mathbf{m},\mathbf{p}}$ with respect to the norm $\|\cdot\|_{2,\infty}$ is known in the literature to be upper bounded as follows [87]:

$$\mathcal{N}_{2,\infty}(\mathcal{K}^{N}_{\mathbf{m},\mathbf{p}},\epsilon) \leq (3/\epsilon)^{\sum_{i=1}^{N} m_i p_i}. \tag{3.16}$$

We now turn to finding the covering numbers of LSR sets $\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}$ and ${}^{c}\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$. The following lemma establishes a bound on covering number of $\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}$, which depends on

the separation rank r exponentially.

Lemma 7. *The covering number of the set $\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}$ with respect to the norm $\|\cdot\|_{2,\infty}$ is upper bounded as follows:*

$$\mathcal{N}_{2,\infty}(\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}},\epsilon) \leq (9p/\epsilon)^{r(m_1p_1+m_2p_2+1)}.$$

Proof. Let $\mathcal{M}^r_{m\times p}$ be the manifold of rank-r matrices on the Euclidean unit ball

$$\mathcal{M}^r_{m\times p} = \{\mathbf{D}\in\mathcal{U}|\,\mathrm{rank}(\mathbf{D})\leq r\}.$$

Moreover, define $\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}} = \mathcal{L}^{2,r}_{\mathbf{m},\mathbf{p}}\cap\mathcal{U}$. Since the rearrangement operator is an isometry w.r.t. the Euclidean distance, the image of an ϵ-net of $\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}}$ w.r.t. the Frobenius norm under this rearrangement operator is an ϵ-net of $\mathcal{M}^r_{m'\times p'}$ ($m'=m_2p_2$ and $p'=m_1p_1$) w.r.t the Frobenius norm. Thus,

$$\mathcal{N}_F(\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}},\epsilon) = \mathcal{N}_F(\mathcal{M}^r_{m'\times p'},\epsilon).$$

Also, from $\mathcal{N}_F(\mathcal{M}^r_{m'\times p'},\epsilon)\leq(9/\epsilon)^{r(m'+p'+1)}$ [46] we have that

$$\mathcal{N}_F(\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}},\epsilon) \leq (9/\epsilon)^{r(m_1p_1+m_2p_2+1)}. \tag{3.17}$$

On the other hand, for the oblique manifold we have $\mathcal{D}_{m\times p}\subset p\mathcal{U}$ and therefore, $\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}\subset p\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}}$. Hence,

$$\mathcal{N}_{2,\infty}(\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}) \leq \mathcal{N}_{2,\infty}(p\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}},\epsilon).$$

Also, since $\|\mathbf{M}\|_{2,\infty}\leq\|\mathbf{M}\|_F$ for any $\mathbf{M}$, it follows that an ϵ-covering of any given set w.r.t. the Frobenius norm is also an ϵ-covering of that set w.r.t. the max-column-norm. Thus

$$\mathcal{N}_{2,\infty}(\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}) \leq \mathcal{N}_{2,\infty}(p\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}},\epsilon) \leq \mathcal{N}_F(p\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}},\epsilon).$$

Moreover, it follows from the fact $\mathcal{N}_F(p\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}}, \epsilon) = \mathcal{N}_F(\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}}, \epsilon/p)$ that

$$\mathcal{N}_{2,\infty}(\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}, \epsilon) \leq \mathcal{N}_F(\widehat{\mathcal{L}}^{2,r}_{\mathbf{m},\mathbf{p}}, \epsilon/p). \tag{3.18}$$

Thus, from (3.17) and (3.18) we see that

$$\mathcal{N}_{2,\infty}(\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}, \epsilon) \leq (9p/\epsilon)^{r(m_1p_1+m_2p_2+1)},$$

which concludes the proof. □

Next, we obtain an upper bound on the covering number of ${}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ for a given constant c.

Lemma 8. *The covering number of the set ${}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ with respect to the max-column norm $\|\cdot\|_{2,\infty}$ is bounded as follows:*

$$\mathcal{N}_{2,\infty}({}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}, \epsilon) \leq (3rc/\epsilon)^{r\sum_{i=1}^N m_ip_i}.$$

Proof. Each element $\mathbf{D} \in {}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ can be written as a summation of at most r KS matrices $\bigotimes \mathbf{D}^k_n$ such that $\left\|\bigotimes \mathbf{D}^k_n\right\|_F \leq c$. This implies that ${}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$ is a subset of the Minkowski sum (vector sum) of r copies of ${}^c\mathcal{K}^{N,1}_{\mathbf{m},\mathbf{p}}$, the set of KS matrices within the Euclidean ball of radius c. It is easy to show that the Minkowski sum of the ϵ-coverings of r sets is an $r\epsilon$-covering of the Minkowski sum of those sets in any norm. Therefore, we have

$$\mathcal{N}_{2,\infty}({}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}, \epsilon) \leq \left(\mathcal{N}_{2,\infty}({}^c\mathcal{K}^{N,1}_{\mathbf{m},\mathbf{p}}, \epsilon/r)\right)^r. \tag{3.19}$$

Moreover, we have ${}^c\mathcal{K}^{N,1}_{\mathbf{m},\mathbf{p}} \subset c\mathcal{K}^N_{\mathbf{m},\mathbf{p}}$. We also know from equation (16) that $\mathcal{N}(\mathcal{K}^N_{\mathbf{m},\mathbf{p}}, \epsilon) \leq (3/\epsilon)^{\sum_{i=1}^N m_ip_i}$. Thus,

$$\begin{aligned}\mathcal{N}_{2,\infty}({}^c\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}, \epsilon) &\leq \left(\mathcal{N}_{2,\infty}(c\mathcal{K}^N_{\mathbf{m},\mathbf{p}}, \epsilon/r)\right)^r \\ &\leq (3rc/\epsilon)^{r\sum_{i=1}^N m_ip_i}. \end{aligned} \tag{3.20}$$

□

Now that we established covering numbers for our constraint sets of interest, we can now find the sample complexity of the LSR-DL Problem (3.4) by plugging in the values of ν and c_0 in Lemma 6.

Theorem 5. *Consider the empirical LSR dictionary learning Problem* (3.4) *with constraint set* $\mathcal{C}$ *being* $\mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}$, $\mathcal{K}^{N}_{\mathbf{m},\mathbf{p}}$, *or* ${}^{c}\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$. *Fix any* $u > 0$. *Suppose the generating dictionary* $\mathbf{D}^0 \in \mathcal{C}$ *and Assumptions 1–3 are satisfied. Assume* $\bar{\lambda}C_{\min} < \rho < \bar{\lambda}C_{\max}$ *and* $\frac{M_\epsilon}{M_x} < \frac{7}{2}(\bar{\lambda}C_{max} - \rho)$. *Define a constant* ν *that depends on the dictionary class:*

- $\nu = \sum_{i=1}^{N} m_i p_i$ *and* $c_0 = 3$ *when* $\mathcal{C} = \mathcal{K}^{N}_{\mathbf{m},\mathbf{p}}$,
- $\nu = 2r(m_1 p_1 + m_2 p_2 + 1)$ *and* $c_0 = 9p$ *when* $\mathcal{C} = \mathcal{K}^{2,r}_{\mathbf{m},\mathbf{p}}$,
- $\nu = r\sum_{i=1}^{N} m_i p_i$ *and* $c_0 = rc$ *when* $\mathcal{C} = {}^{c}\mathcal{K}^{N,r}_{\mathbf{m},\mathbf{p}}$.

Then, given a number of samples L *satisfying*

$$\frac{L}{\log L} \geq Cp^2 \left(\nu \log c_0 + u\right) \frac{M_y^4}{\left(\rho\left(\rho - \bar{\lambda}C_{\min}\right) s\mathbb{E}\{x^2\}\right)^2} \tag{3.21}$$

where C *is a constant, with probability no less than* $1 - e^{-u}$, *the empirical risk objective function* $\mathbf{D} \in \mathcal{C} \mapsto F_{\mathbf{Y}}(\mathbf{D})$ *has a local minimizer* $\mathbf{D}^*$ *such that* $\left\|\mathbf{D}^* - \mathbf{D}^0\right\|_F < \rho$.

Proof. We take a similar approach to the proof of Theorem 4. Due to compactness of the ball $\mathcal{B}_\rho = \{\mathbf{D} \in \mathcal{C} | \left\|\mathbf{D} - \mathbf{D}^0\right\|_F \leq \rho\}$ and continuity of $F_{\mathbf{Y}}(\mathbf{D})$, it follows from the extreme value theorem that $\mathbf{D} \in \mathcal{B}_\rho \mapsto F_{\mathbf{Y}}(\mathbf{D})$ attains its minimum at a point in $\mathcal{B}_\rho$. It remains to show that $\Delta F_{\mathbf{Y}}(\mathbf{D}; \mathbf{D}^0) > 0$ for all $\mathbf{D} \in \partial\mathcal{B}_\rho$ which implies existence of a local minimizer of $F_{\mathbf{Y}} : \mathcal{C} \to \mathbb{R}$ at $\mathbf{D}^*$ such that $\left\|\mathbf{D}^* - \mathbf{D}^0\right\|_F < \rho$.

Inequality (3.15) shows that it suffices to set $\eta_L \leq \Delta f_{\mathcal{P}}(\mathbf{D}; \mathbf{D}^0)/2$ in order to have $\Delta F_{\mathbf{Y}}(\mathbf{D}; \mathbf{D}^0) > 0$. From Lemma 6 we know $\eta_L \geq 3c_1\sqrt{\frac{c_2\nu \log L}{L}} + c_1\sqrt{\frac{c_2\nu + u}{L}}$. Therefore, using the lower bound (3.13) on $\Delta f_{\mathcal{P}}(\rho)$ we have with probability at least $1 - e^{-u}$

$$3c_1\sqrt{\frac{c_2\nu \log L}{L}} + c_1\sqrt{\frac{c_2\nu + u}{L}} \leq \frac{\mathbb{E}\{x^2\}}{16} \cdot \frac{s}{p} \cdot \rho\left(\rho - \bar{\lambda}C_{\min}\right)$$

with $c_1 \geq M_y^2/\sqrt{8}$ and $c_2 \geq \max(1, \log c_0\sqrt{8}M_y)$[10]. Rearranging, we get

$$\frac{L}{\log L} \geq c_1^2 \left(\frac{3\sqrt{c_2\nu} + \sqrt{c_2\nu + u}}{\rho\left(\rho - \bar{\lambda}C_{\min}\right)} \right)^2 \left(\frac{16}{\mathbb{E}\{x^2\}} \cdot \frac{p}{s} \right)^2. \tag{3.22}$$

Setting $c_1 \geq M_y^2/\sqrt{8}$ and $c_2 = c_3 \log c_0 \geq \max(1, \log c_0\sqrt{8}M_y)$ we get the lower bound

$$\frac{L}{\log L} \geq Cp^2 \left(\nu \log c_0 + u\right) \left(\frac{M_y^2}{\rho\left(\rho - \bar{\lambda}C_{\min}\right) s\mathbb{E}\{x^2\}} \right)^2$$

with probability at least $1 - e^{-u}$. Given that the number of samples satisfies (3.21) for $\bar{\lambda}C_{\min} < \rho < \bar{\lambda}C_{\max}$, with high probability $\Delta F_{\mathbf{Y}} > 0$ for any $\mathbf{D} \in \partial\mathcal{B}_\rho$. Therefore, it follows from the existence of the infimum of $\mathbf{D} \in \mathcal{B}_\rho \mapsto F_{\mathbf{Y}}(\mathbf{D})$ in $\mathcal{B}_\rho$ that $\mathbf{D} \in \mathcal{C} \mapsto F_{\mathbf{Y}}(\mathbf{D})$ has a local minimum at a point within a ball of radius ρ around the true dictionary $\mathbf{D}^0$. □

The $\Omega\big(r(\sum_n m_n p_n)p^2\rho^{-2}\big)$ sample complexity we obtain here for rank-constrained LSR-DL is a reduction compared to the $\Omega(mp^3\rho^{-2})$ sample complexity of standard DL in [77]. However, a minimax lower bound scaling of $\Omega(p\sum_n m_n p_n \rho^{-2})$ in [64] for KS-DL ($r = 1$) suggests an $O(p)$ gap with our upper bound.

3.4 Identifiability in the Tractable LSR-DL Problems

In Section 3.2, we introduced two tractable relaxations to the rank-constrained LSR-DL problem: a regularized problem (3.8) with a convex regularization term and a factorization-based problem (3.9) in which the dictionary is written in terms of its subdictionaries. Based on our results in Section 3.3 for the rank-constrained problem, we now provide results on the local identifiability of the true dictionary $\mathbf{D}^0$ in these problems, i.e., we find conditions under which at least one local minimizer of these problems is located near the true dictionary $\mathbf{D}^0$. Such local identifiability result implies that any DL algorithm that converges to a local minimum of these problems can recover

[10] Under the conditions of this theorem, $M_y \leq \sqrt{1 + \delta_s(\mathbf{D}^0)}M_x + M_\epsilon$, where $\delta_s(\mathbf{D}^0)$ denotes the RIP constant of $\mathbf{D}^0$.

$\mathbf{D}^0$ up to a small error if it is initialized close enough to $\mathbf{D}^0$.

3.4.1 Regularization-based LSR Dictionary Learning

The first tractable LSR-DL problem that we study is the regularized problem (3.8). Exploiting the relation between $\mathfrak{R}^N_{\mathbf{m},\mathbf{p}}(\mathbf{D})$ and $\text{rank}(\underline{\mathbf{D}}^\pi)$, the LSR structure is enforced on the dictionary by a convex regularizer that imposes low tensor rank structure on $\underline{\mathbf{D}}^\pi$. The regularizer that we use here is a commonly used convex proxy for the tensor rank function, the *sum-trace-norm* [88], which is defined as the average of the trace (nuclear) norms of the *unfoldings* of the tensor: $\|\underline{\mathbf{A}}\|_{\text{str}} \triangleq \sum_{n=1}^N \left\|\mathbf{A}^{(n)}\right\|_{\text{tr}}$.

The first question we address is whether the reference dictionary that generates the observations $\{\underline{\mathbf{Y}}_l\}_{l=1}^L$ is identifiable via Problem (3.8). Our local identifiability result here is limited to when $\mathbf{D}^0 \in \mathcal{K}^N_{\mathbf{m},\mathbf{p}}$, i.e. the true dictionary is KS. For such $\mathbf{D}^0$, we show that there is at least one local minimizer $\mathbf{D}^*$ of $F^{\text{reg}}_{\mathbf{Y}}(\mathbf{D})$ under Assumptions 1–3 that is close to $\mathbf{D}^0$.

Theorem 6. *Consider the regularized LSR-DL problem* (3.8)*. Suppose that the generating dictionary* $\mathbf{D}^0 \in \mathcal{K}^N_{\mathbf{m},\mathbf{p}}$ *and Assumptions 1–3 are satisfied. Moreover, let* $\bar{\lambda}C_{\min} < \rho \leq \bar{\lambda}C_{\max}$ *and* $\frac{M_\epsilon}{M_x} < \frac{7}{2}(\bar{\lambda}C_{max} - \rho)$*. Then, the expected risk function* $\mathbf{D} \in \mathcal{D} \mapsto \mathbb{E}[F^{\text{reg}}_{\mathbf{Y}}(\mathbf{D})]$ *has a local minimizer* $\mathbf{D}^*$ *such that* $\left\|\mathbf{D}^* - \mathbf{D}^0\right\|_F \leq \rho$*.*

Moreover, given L *samples such that*

$$L > C_0 p^2(mp+u)\left(\frac{M_x^2}{\mathbb{E}x^2}\cdot\frac{\frac{M_\epsilon}{M_x}+\rho+(\frac{M_\epsilon}{M_x}+\rho)^2}{\rho - C_{\min}\bar{\lambda}}\right)^2, \tag{3.23}$$

where u *and* C *are positive constants, then, we have with probability no less than* $1-e^{-u}$ *that the empirical risk function* $\mathbf{D} \in \mathcal{D} \mapsto F^{\text{reg}}_{\mathbf{Y}}(\mathbf{D})$ *has a local minimum at* $\mathbf{D}^*$ *such that* $\left\|\mathbf{D}^* - \mathbf{D}^0\right\|_F < \rho$*.*

Proof. Consider the ball $\mathcal{B}_\rho = \{\mathbf{D} \in \mathcal{D} | \left\|\mathbf{D} - \mathbf{D}^0\right\|_F \leq \rho\}$. It follows from the extreme value theorem [85] that $\mathbf{D} \in \mathcal{B}_\rho \mapsto F^{\text{reg}}_{\mathbf{Y}}(\mathbf{D})$ attains its minimum at a point in $\mathcal{B}_\rho$. This is based on compactness of $\mathcal{B}_\rho = \{\mathbf{D} \in \mathcal{C} | \left\|\mathbf{D} - \mathbf{D}^0\right\|_F \leq \rho\}$ and continuity of $F^{\text{reg}}_{\mathbf{Y}}(\mathbf{D})$. Similarly, $\mathbf{D} \in \mathcal{B}_\rho \mapsto f^{\text{reg}}_{\mathcal{P}}(\mathbf{D}) \triangleq \mathbb{E}[F^{\text{reg}}_{\mathbf{Y}}]$ reaches its minimum at a point in $\mathcal{B}_\rho$. We now

need to show in either case the minimum is not attained on the boundary of $\mathcal{B}_\rho$. To this end, we show in the following that $\Delta F_{\mathbf{Y}}^{\text{reg}}(\mathbf{D};\mathbf{D}^0) > 0$ and $\Delta f_{\mathcal{P}}^{\text{reg}}(\mathbf{D};\mathbf{D}^0) > 0$ for any $\mathbf{D} \in \partial\mathcal{B}_\rho$.

Incorporation of the trace-norm regularization term in (3.8) whithin the objective in (3.4) introduces a factor $\|\underline{\mathbf{D}}^\pi\|_{\text{str}} - \left\|[\underline{\mathbf{D}}^0]^\pi\right\|_{\text{str}} = \sum_{n=1}^N \left(\left\|\mathbf{D}^{(n)}\right\|_{\text{tr}} - \left\|[\mathbf{D}^0]^{(n)}\right\|_{\text{tr}}\right)$ to $\Delta f_{\mathcal{P}}(\mathbf{D};\mathbf{D}^0)$ and $\Delta F_{\mathbf{Y}}(\mathbf{D};\mathbf{D}^0)$. We know from Lemma 2 that when the true dictionary is a KS matrix ($\mathbf{D}^0 \in \mathcal{K}_{\mathbf{m},\mathbf{p}}^N$), its rearrangement tensor $[\underline{\mathbf{D}}^0]^\pi$ is a rank-1 tensor and therefore all unfoldings $[\mathbf{D}^0]^{(n)}$ of $[\underline{\mathbf{D}}^0]^\pi$ are rank-1 matrices. This implies $\|\mathbf{D}^{(n)}\|_{\text{tr}} = \|\mathbf{D}^{(n)}\|_F$. Likewise, for all $\mathbf{D} \in \mathcal{D}_{m\times p}$ we have $\left\|\mathbf{D}^{(n)}\right\|_F = \left\|[\mathbf{D}^0]^{(n)}\right\|_F = \sqrt{p}$. Therefore,

$$\begin{aligned}\left\|\mathbf{D}^{(n)}\right\|_{\text{tr}} - \left\|[\mathbf{D}^0]^{(n)}\right\|_{\text{tr}} &= \sum\nolimits_{k=1}^{r_n} \sigma_k(\mathbf{D}^{(n)}) - \sqrt{p} \\ &\geq \sqrt{\sum\nolimits_{k=1}^{r_n} \sigma_k^2(\mathbf{D}^{(n)})} - \sqrt{p} = 0,\end{aligned}$$

where $r_n \triangleq \text{rank}(\mathbf{D}^{(n)})$ and $\sigma_k(\mathbf{D}^{(n)})$ denotes the k-th singular value of $\mathbf{D}^{(n)}$. Therefore, we conclude that $\Delta F_{\mathbf{Y}}^{\text{reg}}(\mathbf{D};\mathbf{D}^0) \geq \Delta F_{\mathbf{Y}}(\mathbf{D};\mathbf{D}^0)$ and $\Delta f_{\mathcal{P}}^{\text{reg}}(\mathbf{D};\mathbf{D}^0) \geq \Delta f_{\mathcal{P}}(\mathbf{D};\mathbf{D}^0)$ for any $\mathbf{D} \in \mathcal{D}$. According to Lemma 5, $\Delta f_{\mathcal{P}}(\mathbf{D};\mathbf{D}^0) > 0$ for all $\mathbf{D}$ on the boundary of the ball $\mathcal{B}_\rho$. Furthermore, under the assumptions of the current theorem, given a number of samples satisfying (3.23), Gribonval et al. [77] show that the empirical difference $\Delta F_{\mathbf{Y}}(\mathbf{D};\mathbf{D}^0) > 0$ for all $\mathbf{D}$ on the boundary of $\mathcal{S}_\rho = \left\{\mathbf{D} \in \mathcal{D} \;\middle|\; \left\|\mathbf{D} - \mathbf{D}^0\right\|_F \leq \rho\right\}$, and therefore on the boundary of $\mathcal{B}_\rho \subseteq \mathcal{S}_\rho$, with probability at least $1 - e^{-u}$. Therefore, for both $f_{\mathcal{P}}^{\text{reg}}(\mathbf{D})$ and $F_{\mathbf{Y}}^{\text{reg}}(\mathbf{D})$, the minimum is attained in the interior of $\mathcal{B}_\rho$ and not on its boundary. □

3.4.2 Factorization-based LSR Dictionary Learning

We now shift our focus to Problem (3.9), which expands $\mathbf{D}$ as $\sum_{k=1}^r \bigotimes \mathbf{D}_n^k$ and optimizes over the individual subdictionaries, and show that there is at least one local minimum $\{[\mathbf{D}_n^k]^*\}$ of the factorization-based LSR-DL Problem (3.9) such that $\sum \bigotimes [\mathbf{D}_n^k]^*$ is close to the underlying dictionary $\mathbf{D}^0$. Our strategy here is to establish a connection between the local minima of (3.9) and those of (3.4). Specifically, we show that when the dictionary class in (3.7) matches that of (3.9), for every local minimum $\widehat{\mathbf{D}}$ of (3.4),

there exists a local minimum $\{\widehat{\mathbf{D}}_n^k\}$ of (3.9) such that $\widehat{\mathbf{D}} = \sum \bigotimes \widehat{\mathbf{D}}_n^k$. Furthermore, we use the result of Theorems 4 and 5 that there exists a local minimum $\mathbf{D}^*$ of Problem (3.4) within a small ball around $\mathbf{D}^0$. It follows from these facts that under the generating model considered here, a local minimum $\{[\mathbf{D}_n^k]^*\}$ of (3.9) is such that $\sum \bigotimes [\mathbf{D}_n^k]^*$ is close to $\mathbf{D}^0$.

We begin with a bound on the distance between LSR matrices when the tuples of their factor matrices are ϵ-close.

Lemma 9. *For any two tuples* $(\mathbf{A}_n^k)$ *and* $(\mathbf{B}_n^k)$ *such that* $\mathbf{A}_n^k, \mathbf{B}_n^k \in \alpha\mathcal{U}_{m_n \times p_n}$ *for all* $n \in [N]$ *and* $k \in [r]$*, if the distance* $\big\|(\mathbf{A}_n^k) - (\mathbf{B}_n^k)\big\|_F \leq \epsilon$ *then* $\big\| \sum_{k=1}^r \bigotimes \mathbf{A}_n^k - \sum_{k=1}^r \bigotimes \mathbf{B}_n^k \big\|_F \leq \alpha^{N-1}\sqrt{Nr}\epsilon$.

Proof. According to Lemma 2 in Shakeri et al. [17], for any $\{\mathbf{A}_n\}$ and $\{\mathbf{B}_n\}$ we have

$$\begin{aligned} &\bigotimes_{n=1}^N \mathbf{A}_n - \bigotimes_{n=1}^N \mathbf{B}_n \\ &\quad = \sum_{n=1}^N \mathbf{\Gamma}_1 \otimes \cdots \otimes (\mathbf{A}_n - \mathbf{B}_n) \otimes \cdots \otimes \mathbf{\Gamma}_N, \end{aligned} \tag{3.24}$$

where $\mathbf{\Gamma}_n = \mathbf{A}_n$ or $\mathbf{\Gamma}_n = \mathbf{B}_n$ depending on n. Let $\epsilon_n^k \triangleq \|\mathbf{A}_n^k - \mathbf{B}_n^k\|_F$. Using equality (3.24), we have

$$\begin{aligned} &\big\| \sum_{k=1}^r \bigotimes \mathbf{A}_n^k - \sum_{k=1}^r \bigotimes \mathbf{B}_n^k \big\|_F \\ &= \big\| \sum_{k=1}^r \sum_{n=1}^N \mathbf{\Gamma}_1^k \otimes \cdots \otimes (\mathbf{A}_n^k - \mathbf{B}_n^k) \otimes \cdots \otimes \mathbf{\Gamma}_N^k \big\|_F \\ &\leq \sum_{k=1}^r \sum_{n=1}^N \big\| \mathbf{\Gamma}_1^k \otimes \cdots \otimes (\mathbf{A}_n^k - \mathbf{B}_n^k) \otimes \cdots \otimes \mathbf{\Gamma}_N^k \big\|_F \\ &= \alpha^{N-1} \sum_{k=1}^r \sum_{n=1}^N \epsilon_n^k \overset{(a)}{\leq} \alpha^{N-1}\sqrt{Nr}\epsilon, \end{aligned} \tag{3.25}$$

where the inequality (a) follows from $\|(\epsilon_n^k)\|_1 \leq \sqrt{Nr}\ \|(\epsilon_n^k)\|_2 \leq \sqrt{Nr}\epsilon$. □

Theorem 7. *Consider the factorization-based LSR-DL problem* (3.9)*. Suppose that Assumptions 1–3 are satisfied and* $\frac{M_\epsilon}{M_x} < \frac{7}{2}(\bar{\lambda}C_{max} - \rho)$ *with* $\bar{\lambda}C_{\min} < \rho \leq \bar{\lambda}C_{\max}$*. Then, the expected risk function* $\mathbb{E}[F_{\mathbf{Y}}^{\text{fac}}\big(\{\mathbf{D}_n^k\}\big)]$ *has a local minimizer* $([\mathbf{D}_n^k]^*)$ *such that*

$\left\|\sum\bigotimes[\mathbf{D}_n^k]^* - \mathbf{D}^0\right\|_F \leq \rho$.

Moreover, when the sample complexity requirements (3.21) *are satisfied for some positive constant* u*, then with probability no less than* $1-e^{-u}$ *the empirical risk function* $F_{\mathbf{Y}}^{\text{fac}}\big(\{\mathbf{D}_n^k\}\big)$ *has a local minimum achieved at* $([\mathbf{D}_n^k]^*)$ *such that* $\left\|\sum\bigotimes[\mathbf{D}_n^k]^* - \mathbf{D}^0\right\|_F \leq \rho$.

Proof. Let us first consider the finite sample case. Theorem 5 shows existence of a local minimizer $\mathbf{D}^*$ of Problem (3.7) for constraint sets $\mathcal{K}_{\mathbf{m},\mathbf{p}}^N$, $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{2,r}$, and ${}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$, such that $\|\mathbf{D}^* - \mathbf{D}^0\|_F \leq \rho$ w.h.p. Here, we want to show that for such $\mathbf{D}^*$, there exists a $\{[\mathbf{D}_n^k]^*\}$ such that $\mathbf{D}^* = \sum\bigotimes[\mathbf{D}_n^k]^*$ and $\{[\mathbf{D}_n^k]^*\}$ is a local minimizer of Problem (3.9).

First, let us consider Problem (3.7) with ${}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$. It is easy to show that any $\mathbf{D} \in {}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ can be written as $\sum_{k=1}^r\bigotimes\mathbf{D}_n^k$ for all $k \in [r]$ and $n \in [N]$ such that, without loss of generality, $\mathbf{D}_n^k \in \alpha\mathcal{U}_{m\times p}$ where $\alpha = \sqrt[N-1]{c}$. Define

$$\mathcal{C}^{\text{fac}} \triangleq \left\{(\mathbf{D}_n^k) \middle| \sum\bigotimes\mathbf{D}_n^k \in {}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r} : \forall k, n, \mathbf{D}_n^k \in \alpha\mathcal{U}_{m\times p}\right\}.$$

Since $\mathbf{D}^* \in {}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$, there is a $([\mathbf{D}_n^k]^*) \in \mathcal{C}^{\text{fac}}$ such that $\mathbf{D}^* = \sum\bigotimes[\mathbf{D}_n^k]^*$. According to Lemma 9, for any $\{\mathbf{D}_n^k\} \in \mathcal{C}^{\text{fac}}$ it follows from $\left\|(\mathbf{D}_n^k) - ([\mathbf{D}_n^k]^*)\right\|_F \leq \epsilon'$ that $\left\|\sum\bigotimes\mathbf{D}_n^k - \sum\bigotimes[\mathbf{D}_n^k]^*\right\|_F \leq \alpha^{N-1}\sqrt{Nr}\epsilon' = c\sqrt{Nr}\epsilon'$. Since $\mathbf{D}^*$ is a local minimizer of (3.7), there exists a positive ϵ such that for all $\mathbf{D} \in {}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$ satisfying $\|\mathbf{D} - \mathbf{D}^*\|_F \leq \epsilon$, we have $F_{\mathbf{Y}}(\mathbf{D}^*) \leq F_{\mathbf{Y}}(\mathbf{D})$. If we choose ϵ' small enough such that $c\sqrt{Nr}\epsilon' \leq \epsilon$, then for any $(\mathbf{D}_n^k) \in \mathcal{C}^{\text{fac}}$ such that $\left\|(\mathbf{D}_n^k) - ([\mathbf{D}_n^k]^*)\right\|_F \leq \epsilon'$, we have $\left\|\sum\bigotimes\mathbf{D}_n^k - \mathbf{D}^*\right\|_F \leq \epsilon$ and this means that $F_{\mathbf{Y}}^{\text{fac}}\big(\{\mathbf{D}_n^k\}\big) - F_{\mathbf{Y}}^{\text{fac}}\big(\{[\mathbf{D}_n^k]^*\}\big) = F_{\mathbf{Y}}(\sum\bigotimes\mathbf{D}_n^k) - F_{\mathbf{Y}}(\mathbf{D}^*) \geq 0$. Therefore, $([\mathbf{D}_n^k]^*)$ is a local minimizer of Problem (3.9). This concludes our proof for the finite sample case with constraint set ${}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$.

Note that we can write $\mathcal{K}_{\mathbf{m},\mathbf{p}}^N = {}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,1}$ and $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{2,r} = {}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{2,r}$ with $c \geq p$. Therefore, the above results also hold for $\mathcal{K}_{\mathbf{m},\mathbf{p}}^N$ and $\mathcal{K}_{\mathbf{m},\mathbf{p}}^{2,r}$ since they are special cases of ${}^c\mathcal{K}_{\mathbf{m},\mathbf{p}}^{N,r}$.

It is easy to see similar relation exists between the local minima of $f_{\mathbf{y}}(\mathbf{D})$ and $f_{\mathbf{y}}^{\text{fac}}(\{\mathbf{D}_n^k\}) \triangleq \mathbb{E}[F_{\mathbf{Y}}^{\text{fac}}(\{\mathbf{D}_n^k\})]$, proving the asymptotic result in the statement of this theorem. □

3.4.3 Discussion

In this section, we discuss the local identifiability of the true dictionary in the regularization based formulation and the factorization-based formulation. For the regularization-based formulation, our results only hold for the case where the true dictionary is KS, i.e. $\mathbf{D} \in \mathcal{K}^N_{\mathbf{m},\mathbf{p}}$. We obtain sample complexity requirement of $\Omega(mp^3\rho^{-2})$ in this case, which matches the sample complexity requirement of the unstructured formulation [77]. We believe there is room to improve this result as future work.

For the factorization-based formulation, we show that $\Omega(p\rho^{-2}r\sum_n m_n p_n)$ samples are required for local identifiability of a dictionary of separation-rank r. This result matches that of our intractable formulation. Note that when the separation rank is 1, this result gives a bound on the sample complexity of the KS-DL model as a special case. Unlike the analysis in [17] (limited to KS-DL model) where they obtain a sample complexity of $L = \max_{n\in\{1,\dots,N\}} \Omega(m_n p_n^3 \rho_n^{-2})$, our analysis of the factorized model does not ensure identifiability of the true subdictionaries in the LSR-DL model. However, the result in [17] requires the dictionary coefficient vectors to follow the separable sparsity model. In contrast, our result does not require any constraints on the sparsity pattern of the coefficient vector.

3.5 Computational Algorithms

In Section 3.4, we showed that the tractable LSR-DL Problems (3.8) and (3.9) each have at least one local minimum close to the true dictionary. In this section we develop algorithms to find these local minima. Solving Problems (3.8) and (3.9) require minimization with respect to (w.r.t.) $\mathbf{X} \triangleq [\mathbf{x}_1^T, \cdots, \mathbf{x}_L^T]$. 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).

The sparse coding stage is a simple Lasso problem [89, 90] and remains the same in our algorithms. However, the algorithms differ in their dictionary update stages, which we discuss next.

Remark. We leave the formal convergence results of our algorithms to future work. However, we provide a discussion on challenges and possible approaches to establish convergence of our algorithms in Appendix A, Section 3.5.4.

3.5.1 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 (3.8) and discussed in Section 3.4 using the Alternating Direction Method of Multipliers (ADMM) [91].

The main novelty in solving (3.8) using $g_1(\underline{\mathbf{D}}^\pi) = \|\underline{\mathbf{D}}^\pi\|_{\text{str}}$ 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 [88, 92, 93], we instead suggest the following reformulation of the dictionary update stage of (3.8):

$$\begin{aligned} \min_{\mathbf{D}\in\mathcal{D},\underline{\mathbf{W}}_1,\cdots,\underline{\mathbf{W}}_N} \quad & \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \lambda_1 \sum_{n=1}^{N} \left\|\mathbf{W}_n^{(n)}\right\|_{\text{tr}} \\ \text{s.t.} \quad & \forall n \quad \underline{\mathbf{W}}_n = \underline{\mathbf{D}}^\pi. \end{aligned} \tag{3.26}$$

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 (3.26) (without $\mathbf{D} \in \mathcal{D}$ constraint) using ADMM, which involves decoupling the problem into independent subproblems by forming the following augmented Lagrangian:

$$\mathcal{L}_\gamma = \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \sum_{n=1}^{N}\Big(\lambda_1 \left\|\mathbf{W}_n^{(n)}\right\|_{\text{tr}} - \langle \underline{\mathbf{A}}_n,\ \underline{\mathbf{D}}^\pi - \underline{\mathbf{W}}_n\rangle + \frac{\gamma}{2}\|\ \underline{\mathbf{D}}^\pi - \underline{\mathbf{W}}_n\|_F^2\Big), \tag{3.27}$$

where $\mathcal{L}_\gamma$ is shorthand for $\mathcal{L}_\gamma(\underline{\mathbf{D}}^\pi, \{\underline{\mathbf{W}}_n\}, \{\underline{\mathbf{A}}_n\})$. In order to find the gradient of (3.27) with respect to $\underline{\mathbf{D}}^\pi$, we rewrite the Lagrangian function in the following form:

$$\mathcal{L}_\gamma = \frac{1}{2}\left\|\mathbf{y} - \mathcal{T}(\underline{\mathbf{D}}^\pi)\right\|_2^2 + \sum_{n=1}^{N}\Big(\lambda_1 \left\|\mathbf{W}_n^{(n)}\right\|_{\text{tr}} - \langle \underline{\mathbf{A}}_n,\ \underline{\mathbf{D}}^\pi - \underline{\mathbf{W}}_n\rangle + \frac{\gamma}{2}\left\|\ \underline{\mathbf{D}}^\pi - \underline{\mathbf{W}}_n\right\|_F^2\Big).$$

Here, $\mathbf{y} \triangleq \text{vec}(\mathbf{Y})$ (not to be confused with our earlier use of $\mathbf{y}$ for $\text{vec}(\underline{\mathbf{Y}})$) and the linear operator $\mathcal{T}(\underline{\mathbf{D}}^\pi) \triangleq \text{vec}(\mathbf{DX}) = \widetilde{\mathbf{X}}^T\mathbf{\Pi}^T\,\text{vec}(\underline{\mathbf{D}}^\pi)$, where $\widetilde{\mathbf{X}} = \mathbf{X} \otimes \mathbf{I}_m$ and $\mathbf{\Pi}$ is a permutation matrix such that $\text{vec}(\underline{\mathbf{D}}^\pi) = \mathbf{\Pi}\,\text{vec}(\mathbf{D})$. The procedure to find $\mathbf{\Pi}$ is explained in Appendix A, Section 6.1. 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 [91]:

$$\underline{\mathbf{D}}^\pi(\tau) = \underset{\underline{\mathbf{D}}^\pi}{\text{argmin}}\,\mathcal{L}_\gamma(\underline{\mathbf{D}}^\pi, \underline{\mathbf{W}}_n(\tau-1), \underline{\mathbf{A}}_n(\tau-1)), \tag{3.28}$$

$$\underline{\mathbf{W}}_n(\tau) = \underset{\underline{\mathbf{W}}_n}{\text{argmin}}\,\mathcal{L}_\gamma(\underline{\mathbf{D}}^\pi(\tau), \underline{\mathbf{W}}_n, \underline{\mathbf{A}}_n(\tau-1)),\ \forall n \in [N], \tag{3.29}$$

$$\underline{\mathbf{A}}_n(\tau) = \underline{\mathbf{A}}_n(\tau-1) - \gamma\left(\underline{\mathbf{D}}^\pi(\tau) - \underline{\mathbf{W}}_n(\tau)\right),\ \forall\ n \in [N]. \tag{3.30}$$

The solution to (3.28) 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

$$\frac{\partial \mathcal{L}_\gamma}{\partial \underline{\mathbf{D}}^\pi} = \mathcal{T}^*(\mathcal{T}(\underline{\mathbf{D}}^\pi) - \mathbf{y}) - \sum_{n=1}^{N}\underline{\mathbf{A}}_n + \sum_{n=1}^{N}\gamma\left(\underline{\mathbf{D}}^\pi - \underline{\mathbf{W}}_n\right),$$

where $\mathcal{T}^*(\mathbf{v}) = \text{vec}^{-1}\left(\mathbf{\Pi}\widetilde{\mathbf{X}}\mathbf{v}\right)$ is the *adjoint* of the linear operator $\mathcal{T}$[93]. Setting the gradient to zero results in

$$\mathcal{T}^*(\mathcal{T}(\underline{\mathbf{D}}^\pi)) + \gamma N\ \underline{\mathbf{D}}^\pi = \mathcal{T}^*(\mathbf{y}) + \sum_{n=1}^{N}\left(\underline{\mathbf{A}}_n + \gamma\underline{\mathbf{W}}_n\right).$$

Equivalently, we have

$$\text{vec}^{-1}\left(\left[\mathbf{\Pi}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\mathbf{\Pi}^T+\gamma N\mathbf{I}\right]\text{vec}(\underline{\mathbf{D}}^\pi)\right)=\text{vec}^{-1}(\mathbf{\Pi}\widetilde{\mathbf{X}}\mathbf{y})+\sum_{n=1}^{N}(\underline{\mathbf{A}}_n+\gamma\underline{\mathbf{W}}_n)\,. \tag{3.31}$$

Therefore, the update rule for $\underline{\mathbf{D}}^\pi$ is

$$\begin{aligned}\underline{\mathbf{D}}^\pi(\tau)=\text{vec}^{-1}\Big(&\left[\mathbf{\Pi}^T\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\mathbf{\Pi}+\gamma N\mathbf{I}_{mp}\right]^{-1}\\&\cdot\left[\mathbf{\Pi}^T\widetilde{\mathbf{X}}\mathbf{y}+\text{vec}\left(\sum_{n=1}^{N}(\underline{\mathbf{A}}_n(\tau-1)+\gamma\underline{\mathbf{W}}_n(\tau-1))\right)\right]\Big).\end{aligned} \tag{3.32}$$

To update $\{\underline{\mathbf{W}}_n\}$, we can further split (3.29) into N independent subproblems (suppressing the index τ):

$$\min_{\underline{\mathbf{W}}_n}\ \mathcal{L}_{\mathcal{W}}=\lambda_1\left\|\mathbf{W}_n^{(n)}\right\|_{\text{tr}}-\langle\underline{\mathbf{A}}_n,\ \underline{\mathbf{D}}^\pi-\underline{\mathbf{W}}_n\rangle+\frac{\gamma}{2}\parallel\underline{\mathbf{D}}^\pi-\underline{\mathbf{W}}_n\parallel_F^2.$$

We can reformulate $\mathcal{L}_{\mathcal{W}}$ as

$$\lambda_1\left\|\mathbf{W}_n^{(n)}\right\|_{\text{tr}}+\frac{\gamma}{2}\left\|\mathbf{W}_n^{(n)}-\left([\underline{\mathbf{D}}^\pi]^{(n)}-\frac{\mathbf{A}_n^{(n)}}{\gamma}\right)\right\|_F^2+\text{const.}$$

The minimizer of $\mathcal{L}_{\mathcal{W}}$ with respect to $\mathbf{W}_n^{(n)}$ is $\text{shrink}\left([\mathbf{D}^\pi]^{(n)}-\frac{1}{\gamma}\mathbf{A}_n^{(n)},\ \frac{\lambda_1}{\gamma}\right)$ where $\text{shrink}(\mathbf{A},z)$ applies soft thresholding at level z on the singular values of matrix $\mathbf{A}$ [94]. Therefore,

$$\underline{\mathbf{W}}_n(\tau)=\text{refold}\Big(\text{shrink}\left([\mathbf{D}^\pi]^{(n)}(\tau)-\frac{1}{\gamma}\mathbf{A}_n^{(n)}(\tau-1),\ \frac{\lambda_1}{\gamma}\right)\Big), \tag{3.33}$$

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.

3.5.2 TeFDiL: A Factorization-based LSR-DL Algorithm

While our experiments in Section 3.6 validate good performance of STARK, the algorithm finds the dictionary $\mathbf{D}\in\mathbb{R}^{m\times p}$ and not the subdictionaries $\{\mathbf{D}_n\in\mathbb{R}^{m_n\times p_n}\}_{n=1}^N$.

[11] In the body of Algorithms 1–3 we drop the iteration index t for simplicity.

Algorithm 1 Dictionary Update in STARK for LSR-DL

Require: $\mathbf{Y}$, $\mathbf{\Pi}$, $\lambda_1 > 0$, $\gamma > 0$, $\mathbf{X}(t)$[11]

repeat
  Update $\underline{\mathbf{D}}^\pi$ according to update rule (3.32)
  for $n \in [N]$ **do**
    Update $\underline{\mathbf{W}}_n$ according to (3.33)
  end for
  for $n \in [N]$ **do**
    $\underline{\mathbf{A}}_n \leftarrow \underline{\mathbf{A}}_n - \gamma\,(\underline{\mathbf{D}}^\pi - \underline{\mathbf{W}}_n)$
  end for
until convergence
Normalize columns of $\mathbf{D}$
return $\mathbf{D}(t+1)$

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 subdictionaries and allows for direct tuning of the separation rank to control the number of parameters of the model. To this end, we propose a factorization-based LSR-DL algorithm termed *Tensor Factorization-Based DL (TeFDiL)* in this section for solving Problem (3.9).

We discussed earlier in Section 3.5.1 that the error term $\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2$ can be reformulated as $\|\mathbf{y} - \mathcal{T}(\underline{\mathbf{D}}^\pi)\|^2$ where $\mathcal{T}(\underline{\mathbf{D}}^\pi) = \widetilde{\mathbf{X}}^T\mathbf{\Pi}^T \operatorname{vec}(\underline{\mathbf{D}}^\pi)$. Thus, the dictionary update objective in (3.9) can be reformulated as $\|\mathbf{y} - \mathcal{T}(\sum_{k=1}^r \mathbf{d}_N^k \circ \cdots \circ \mathbf{d}_1^k)\|^2$ where $\mathbf{d}_n^k = \operatorname{vec}(\mathbf{D}_n^k)$. To avoid the complexity of solving this problem, we resort to first obtaining an inexact solution by minimizing $\|\mathbf{y} - \mathcal{T}(\underline{\mathbf{A}})\|^2$ over $\underline{\mathbf{A}}$ and then enforcing the low-rank structure by finding the rank-r approximation of the minimizer of $\|\mathbf{y} - \mathcal{T}(\underline{\mathbf{A}})\|^2$. TeFDiL employs CP decomposition (CPD) 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[12], then $\widetilde{\mathbf{X}}^T$ is full column-rank and $\underline{\mathbf{A}} = \mathcal{T}^+(\mathbf{y}) = \operatorname{vec}^{-1}\left(\mathbf{\Pi}\big(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\big)^{-1}\widetilde{\mathbf{X}}\mathbf{y}\right)$ minimizes $\|\mathbf{y} - \mathcal{T}(\underline{\mathbf{A}})\|^2$. Now, it remains to solve the following problem to update $\{\mathbf{d}_n^k\}$:

$$\min_{\{\mathbf{d}_n^k\}} \Big\| \sum_{k=1}^r \mathbf{d}_N^k \circ \cdots \circ \mathbf{d}_1^k - \mathcal{T}^+(\mathbf{y}) \Big\|_F^2.$$

[12]In our experiments, we add $\delta\mathbf{I}$ to $\mathbf{X}\mathbf{X}^T$ with a small $\delta > 0$ at every iteration to ensure full-rankness.

The problem of finding the best rank-r approximation (r-term CPD) of a tensor is ill-posed in general in that a solution may not exist for $r > 1$ and $N > 2$, due to the fact that the set over which one optimizes is not closed [86]. However, various numerical algorithms exist in the tensor recovery literature to find a "good" rank-r approximation of a tensor [71, 86] by updating . Perhaps the most common yet simplest of CP Decomposition algorithms is alternating least squares (ALS).

TeFDiL can employ any CP Decomposition algorithm to find the r-term CPD, denoted by $\text{CPD}_r(\cdot)$, of $\mathcal{T}^+(\mathbf{y})$. At the end of each dictionary update stage, the columns of $\mathbf{D} = \sum \bigotimes \mathbf{D}_n^k$ 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 $\mathcal{T}^+(\mathbf{y}) = \text{vec}^{-1}\left(\mathbf{\Pi}\big(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\big)^{-1}\widetilde{\mathbf{X}}\mathbf{y}\right)$
2: $\underline{\mathbf{D}}^\pi \leftarrow \text{CPD}_r(\mathcal{T}^+(\mathbf{y}))$
3: $\mathbf{D} \leftarrow \text{vec}^{-1}\left(\mathbf{\Pi}^T \text{vec}(\underline{\mathbf{D}}^\pi)\right)$
4: Normalize columns of $\mathbf{D}$
5: **return** $\mathbf{D}(t+1)$

3.5.3 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

$$J_T(\{\mathbf{D}_n^k\}) = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{y}(t) - \big(\sum_{k=1}^{r} \bigotimes_{n=1}^{N} \mathbf{D}_n^k\big)\mathbf{x}(t)\|^2.$$

We can rewrite this objective as

$$\begin{aligned} J_T &= \sum_{t=1}^{T} \|\mathbf{Y}^{(n)}(t) - \sum_{k=1}^{r} \mathbf{D}_n^k \mathbf{X}^{(n)}(t)\mathbf{C}_n^k(t)\|_F^2 \\ &= \sum_{t=1}^{T} \|\widehat{\mathbf{Y}}^{(n)}(t) - \mathbf{D}_n^k \mathbf{X}^{(n)}(t)\mathbf{C}_n^k(t)\|_F^2 \\ &= \operatorname{Tr}\Big([\mathbf{D}_n^k]^T \mathbf{D}_n^k \mathbf{A}_n^k(t)\Big) - 2\operatorname{Tr}\Big([\mathbf{D}_n^k]^T \mathbf{B}_n^k(t)\Big) + \text{const.}, \end{aligned}$$

where, dropping the iteration index t, the matrix

$$\mathbf{C}_n^k \triangleq \Big(\mathbf{D}_N^k \otimes \cdots \otimes \mathbf{D}_{n+1}^k \otimes \mathbf{D}_{n-1}^k \cdots \otimes \mathbf{D}_1^k\Big)^T$$

and the estimate $\widehat{\mathbf{Y}}^{(n)} \triangleq \mathbf{Y}^{(n)} - \sum_{\substack{i=1 \\ i\neq k}}^{r} \mathbf{D}_n^i \mathbf{X}^{(n)} \mathbf{C}_n^i$. We can further define the matrices

$$\mathbf{A}_n^k(t) \triangleq \sum_{\tau=1}^{t} \mathbf{X}^{(n)}(t)\mathbf{C}_n^k(\tau)[\mathbf{C}_n^k(\tau)]^T[\mathbf{X}^{(n)}(\tau)]^T \in \mathbb{R}^{p_n \times p_n}$$

and

$$\mathbf{B}_n^k(t) \triangleq \sum_{\tau=1}^{t} \widehat{\mathbf{Y}}^{(n)}(\tau)[\mathbf{C}_n^k(\tau)]^T[\mathbf{X}^{(n)}(\tau)]^T \in \mathbb{R}^{m_n \times p_n}.$$

To minimize J_T with respect to each $\mathbf{D}_n^k$, we take a similar approach as in Mairal et al. [57] and use a (block) coordinate descent algorithm with warm start to update the columns of $\mathbf{D}_n^k$ in a cyclic manner. Algorithm 3 describes the dictionary update step of OSubDil.

3.5.4 Discussion on Convergence of the Algorithms

The batch algorithms proposed in Section 3.5 are essentially variants of alternating minimization (AM). Establishing the convergence of AM-type algorithms in general is challenging and only known for limited cases. Here, we first present a well-known convergence result for AM-type algorithms in Lemma 10 and discuss why our algorithms

Algorithm 3 Dictionary Update in OSubDil for LSR-DL

Require: $\underline{\mathbf{Y}}(t)$, $\{\mathbf{D}_n^k(t)\}$, $\mathbf{A}_n^k(t)$, $\mathbf{B}_n^k(t)$, $\underline{\mathbf{X}}(t)$
for all $k \in [r]$ **do**
  for all $n \in [N]$ **do**
    $\mathbf{C}_n^k \leftarrow \left(\mathbf{D}_N^k \otimes \cdots \otimes \mathbf{D}_{n+1}^k \otimes \mathbf{D}_{n-1}^k \cdots \otimes \mathbf{D}_1^k\right)^T$
    $\widehat{\mathbf{Y}}^{(n)} \leftarrow \mathbf{Y}^{(n)} - \sum_{\substack{i=1 \\ i \neq k}}^{r} \mathbf{D}_n^i \mathbf{X}^{(n)} \mathbf{C}_n^i$
    $\mathbf{A}_n^k \leftarrow \mathbf{A}_n^k + \mathbf{X}^{(n)} \mathbf{C}_n^k [\mathbf{C}_n^k]^T [\mathbf{X}^{(n)}]^T$
    $\mathbf{B}_n^k \leftarrow \mathbf{B}_n^k + \widehat{\mathbf{Y}}^{(n)} [\mathbf{C}_n^k]^T [\mathbf{X}^{(n)}]^T$
    for $j = 1, \cdots, p_n$ **do**
      $[\mathbf{D}_n^k]_j \leftarrow \frac{1}{[\mathbf{A}_n^k]_{jj}} ([\mathbf{B}_n^k]_j - \mathbf{D}_n^k [\mathbf{A}_n^k]_j) + [\mathbf{D}_n^k]_j$
    end for
  end for
end for
Normalize columns of $\mathbf{D} = \sum_{n=1}^{r} \bigotimes_{n=1}^{N} \mathbf{D}_n^k$
return $\{\mathbf{D}_n^k(t+1)\}$

STARK and TeFDiL do not satisfy the requirements of this lemma. Then, we show a possible approach for proving convergence of STARK. We do not discuss convergence analysis of OSubDil here since it does not fall in the batch AM framework that we discuss here. We leave formal convergence results of our algorithms as open problems for future work.

First, let us state the following standard convergence result for AM-type algorithms.

Lemma 10 (Proposition 2.7.1, [95])**.** *Consider the problem*

$$\min_{\mathbf{x}=(\mathbf{x}_1,\ldots,\mathbf{x}_M) \in \mathcal{E}=\mathcal{E}_1 \times \mathcal{E}_2 \times \cdots \times \mathcal{E}_M} f(\mathbf{x})$$

where $\mathcal{E}_i$ are closed convex subsets of the Euclidean space. Assume that $f(\cdot)$ is a continuous differentiable over the set $\mathcal{E}$. Suppose for each i and all $\mathbf{x} \in \mathcal{E}$, the minimum

$$\min_{\xi \in \mathcal{E}_i} f(\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}, \xi, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_M)$$

is uniquely attained. Then every limit point of the sequence $\{\mathbf{x}(t)\}$ generated by block coordinate descent method is a stationary point of $f(\cdot)$.

The result of Lemma 10 cannot be used for TeFDiL since its dictionary update stage does not have a unique minimizer (nonconvex minimization problem with multiple

global minima)). Moreover, as discussed in Section 3.5.2, TeFDiL only returns an inexact solution.

Similarly, this result cannot be used to show convergence of STARK to a stationary point of Problem (3.8) due to the fact that the constraint set $\mathcal{D}_{m\times p}$ is not convex. However, we show next that dropping the unit column-norm constraint allows us to provide certain convergence guarantees. The unit column-norm constraint is essential in standard DL algorithms since in its absence, the ℓ_1 norm regularization term encourages undesirable solutions where $\|\mathbf{X}\|_F$ is very small while $\|\mathbf{D}\|_F$ is very large. However, in the regularization-based LSR-DL problem, the additional regularization term $\|\underline{\mathbf{D}}^{\pi}\|_{\text{str}}$ ensures this does not happen. Therefore, dropping the unit column-norm constraint is sensible in this problem.

Let us discuss what guarantees we are able to obtain after relaxing the constraint set $\mathcal{D}_{m\times p}$. Consider the minimization problem

$$\min_{\mathbf{D}\in\mathbb{R}^{m\times p},\mathbf{X}\in\mathbb{R}^{p\times L}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \lambda_1 \|\underline{\mathbf{D}}^{\pi}\|_{\text{str}} + \lambda\|\mathbf{X}\|_{1,1}. \tag{3.34}$$

We show that under the following assumptions, STARK converges to a stationary point of Problem (3.34) (when the normalization step is not enforced). Then we discuss how this problem is related to Problem (3.8).

Assumption 4. *Consider the sequence* $\big(\mathbf{D}(t),\mathbf{X}(t)\big)$ *generated by STARK. We assume that for all* $t\geq 0$*:*

I) Classical optimality conditions for the lasso problem (see Tibshirani [96]) are satisfied.

II) $\mathbf{X}(t)$ *is full row-rank at all* t*.*

Proposition 1 establishes the convergence of STARK (without normalization).

Proposition 1. *Under Assumption 4, STARK converges to a stationary point of problem* (3.34)*.*

Proof. We invoke Lemma 10 to show the convergence of STARK. To use this lemma,

the minimization problem w.r.t. each block needs to correspond to a closed convex constraint set and also needs to have a unique minimizer.

In the sparse coding stage, given Assumption 4-I, the minimizer of the lasso problem is unique. In the dictionary update stage of STARK, the objective of problem (3.34) is strongly convex w.r.t. $\mathbf{D}$ under Assumption 4-II and thus has a unique minimizer. Moreover, the constraint set $\mathbb{R}^{p\times L}$ is closed and convex. To utilize Lemma 10, it remains to show that this minimum is actually attained by ADMM. To this end, we restate Problem (3.26) as

$$\begin{aligned} \min_{\underline{\mathbf{D}}^{\pi},\widetilde{\underline{\mathbf{W}}}} \quad & f_1(\underline{\mathbf{D}}^{\pi}) + f_2(\widetilde{\underline{\mathbf{W}}}) \\ \text{s.t.} \quad & \widetilde{\underline{\mathbf{W}}} = \mathcal{H}\underline{\mathbf{D}}^{\pi}, \end{aligned} \tag{3.35}$$

where $f_1(\underline{\mathbf{D}}^{\pi}) = \frac{1}{2}\left\|\mathbf{Y}-\mathbf{D}\mathbf{X}\right\|_F^2$ (note that $\mathbf{D}\mathbf{X}$ is a linear function of $\underline{\mathbf{D}}^{\pi}$) and $f_2(\widetilde{\underline{\mathbf{W}}}) = \lambda_1 \sum_{n=1}^{N} \left\|(\underline{\mathbf{W}}_n)_{(n)}\right\|_*$. It is clear that $\mathcal{H}\mathcal{H}^*$ is convertible. Therefore, according to Lemma 11 stated below, the ADMM algorithm converges to the unique minimizer of Problem (3.26).

Lemma 11 (Chapter 3, Proposition 4.2, [97])**.** *Consider the problem*

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{C}_1,\mathbf{z}\in\mathcal{C}_2} \quad & f_1(\mathbf{x}) + f_2(\mathbf{z}) \\ s.t. \quad & \mathbf{z} = \mathcal{A}(\mathbf{x}) \end{aligned} \tag{3.36}$$

Then, if $\mathcal{A}\mathcal{A}^*$ is invertible or if $\mathcal{C}_1$ is bounded, the sequence generated by the ADMM algorithm applied to the Augmented Lagrangian function converges to an optimum of (3.36)*.*

This concludes the proof. □

So far we discussed convergence of STARK to Problem (3.34) while our identifiability results are for problem (3.8). There is, however, a strong connection between minimization Problems (3.8) and (3.34): for each local minimum $\widehat{\mathbf{D}}$ of problem (3.8), there exists an $\widehat{\mathbf{X}}$ such that $(\widehat{\mathbf{D}},\widehat{\mathbf{X}})$ is a local minimum of (3.34). Define

$\ell_{\mathbf{Y}}^{\text{reg}}(\mathbf{D},\mathbf{X}) = \frac{1}{L}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \lambda_1 \|\underline{\mathbf{D}}^{\pi}\|_{\text{str}} + \frac{\lambda}{L}\|\mathbf{X}\|_{1,1}$. Consider any $\widehat{\mathbf{D}}$ that is a local minimum of (3.8) and let $\widehat{\mathbf{X}} = \text{argmin}_{\mathbf{X}\in\mathbb{R}^{p\times L}} \ell_{\mathbf{Y}}^{\text{reg}}(\widehat{\mathbf{D}},\mathbf{X})$. We have $\ell_{\mathbf{Y}}^{\text{reg}}(\widehat{\mathbf{D}},\widehat{\mathbf{X}}) = F_{\mathbf{Y}}^{\text{reg}}(\widehat{\mathbf{D}})$. Since $\widehat{\mathbf{D}}$ is a local minimizer of $F_{\mathbf{Y}}^{\text{reg}}(\mathbf{D})$, $F_{\mathbf{Y}}^{\text{reg}}(\widehat{\mathbf{D}}) \leq F_{\mathbf{Y}}^{\text{reg}}(\mathbf{D})$ for any $\mathbf{D}$ in the local neighborhood of $\widehat{\mathbf{D}}$. Also by definition, $F_{\mathbf{Y}}^{\text{reg}}(\mathbf{D}) \leq \ell_{\mathbf{Y}}^{\text{reg}}(\mathbf{D},\mathbf{X})$ for any $\mathbf{X}$. Thus, $\ell_{\mathbf{Y}}^{\text{reg}}(\widehat{\mathbf{D}},\widehat{\mathbf{X}}) \leq \ell_{\mathbf{Y}}^{\text{reg}}(\mathbf{D},\mathbf{X})$ for any $(\mathbf{D},\mathbf{X})$ in the local neighborhood of $(\widehat{\mathbf{D}},\widehat{\mathbf{X}})$, meaning that $(\widehat{\mathbf{D}},\widehat{\mathbf{X}})$ is a local minimizer of (3.34). Since we showed in Section 3.4 that a local minimum $\mathbf{D}^*$ of (3.8) is close to the true dictionary $\mathbf{D}^0$, we can say there is a local minimum $(\mathbf{D}^*,\mathbf{X}^*)$ of (3.34) such that $\mathbf{D}^*$ is close to $\mathbf{D}^0$. So our recovery result for (3.8) can apply to our proposed algorithm for solving (3.34) as well.

3.6 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[13] (standard DL) as well as a simple block coordinate descent (BCD) algorithm that alternates between updating every subdictionary in problem (3.9). This BCD algorithm can be interpreted as an extension of the KS-DL algorithm [79] for the LSR model. We show

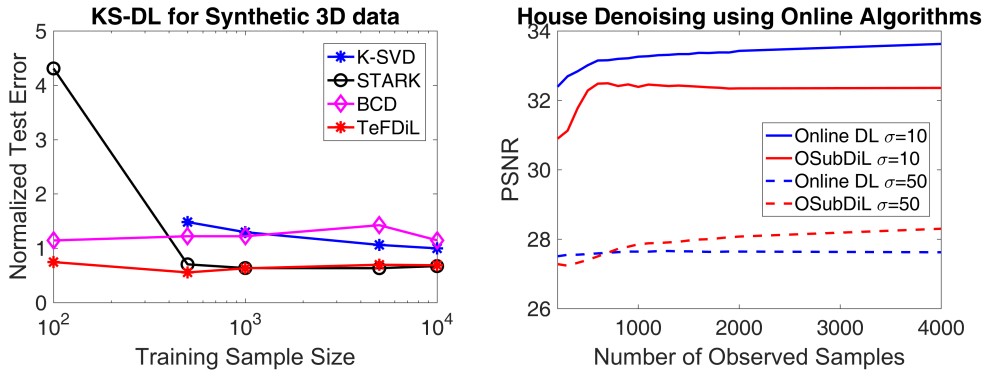
Table 3.2: Performance of DL algorithms for image denoising in terms of PSNR

		Unstructured	KS-DL ($r = 1$)			LSR-DL ($r > 1$)		
Image	Noise	K-SVD [13]	SeDiL [60]	BCD [79]	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 3.3: 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

how structured DL algorithms outperform the unstructured algorithm K-SVD[13] 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 subdictionary elements, and then normalize the columns of the subdictionaries. 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 $\mathbf{y} = \mathbf{D}\mathbf{x}$. 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[98]. 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 3.3a. It can be seen that TeFDiL outperforms all the other algorithms.



(a) (b)

Figure 3.3: (a) Normalized representation error of various DL algorithms for 3rd-order synthetic tensor data. (b) Performance of online DL algorithms for `House`.

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 [99] (to reduce training time) with regularization parameter $\lambda = 0.1$ for all algorithms. To perform denoising, 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 [100]. Table 3.2 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 [60] and BCD [79]. Our LSR-DL methods outperform SeDiL and while BCD has a good performance for $\sigma = 10$, its denoising performance suffers when noise level increases.[13]

Table 3.3 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, 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 subdictionaries and does not allow for direct tuning of the separation rank. Ultimately, the choice between these

[13]Note that SeDiL results may be improved by careful parameter tuning.

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 3.3b shows the PSNR for reconstructing `House` using OSubDil and Online DL in [57] 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.

3.7 Conclusion and Future Work

We studied the low separation rank model (LSR-DL) to learn structured dictionaries for tensor data. This model bridges the gap between unstructured and separable dictionary learning (DL) models. For the intractable rank-constrained and the tractable factorization-based LSR-DL formulations, we show that given $\Omega\big(r(\sum_n m_n p_n)p^2\rho^{-2}\big)$ data samples, the true dictionary can be locally recovered up to distance ρ. This is a reduction compared to the $\Omega(mp^3\rho^{-2})$ sample complexity of standard DL in [77]. However, a minimax lower bound scaling of $\Omega(p\sum_n m_n p_n \rho^{-2})$ in [64] for KS-DL ($r = 1$) has an $O(p)$ gap with our upper bound. One future direction is to close this gap. Furthermore, we show in the regularization-based formulation that $\Omega(mp^3\rho^{-2})$ samples are sufficient for local identifiability of the true Kronecker-structured (KS) dictionary up to distance ρ. Improving this result and providing sample complexity results for when the true dictionary is LSR (and not just KS) is also another interesting future work.

Another interesting theoretical direction of work is providing global identifiability guarantees for the LSR-DL problem. The first hurdle in this direction is that, as mentioned in the introduction of this chapter, our choice of Frobenius norm as the metric results in an optimization problem with multiple global minima, therefore convergence

to a global minimum does not necessarily mean global identifiability. An interesting future direction is to consider alternative (permutation and sign-invariant) distances that result in a single global minimum. The second obstacle in this direction is the difficulty in establishing global convergence results for nonconvex optimization problems. In the recent years, researchers have proposed DL algorithms guaranteed to converge to global optimizers of the nonconvex DL problem [80, 101, 102]. Moreover, Sun et al. [103] show that for the special case of complete dictionary learning, the local minima of the problem are all globally optimum and the saddle points are escapable. While establishing local identifiability is an important first step, obtaining geometric characterization of the optimization landscape of the LSR-DL problem and developing algorithms with global convergence guarantees is an interesting future direction.

Finally, 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 [13] and existing KS-DL algorithms SeDiL [60] and BCD [79]. We also present 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 [57] when the noise level in data is high.

The experimental and theoretical results in this chapter and in the related literature showcase the benefits of exploiting tensor structure of data in the dictionary learning problem. Inspired by these results, in the next chapter we study the benefits of exploiting tensor structure of data in another learning problem, namely the linear regression. Some of the analytical tools used in this chapter will also prove useful in our analysis of tensor linear regression in the next chapter.

Chapter 4

Tensor Regression

In this chapter, we study a tensor-structured linear regression model with tensor-structured predictor and regression parameters and scalar response variables. We focus on the fundamental limits on the accuracy and the sample complexity of estimating the tensor-valued regression parameters (regression tensors) in this model. Specifically, we obtain a lower bound on the minimax risk of estimating the underlying N-th order regression tensor $\underline{\mathbf{B}}^* \in \mathbb{R}^{m_1 \times \cdots \times m_N}$ from L predictor-response pairs $(\underline{\mathbf{X}}_l, y_l)$. By comparing this lower bound to the known lower bounds for standard linear regression, we provide an insight into the benefits of exploiting the tensor structure of $\underline{\mathbf{B}}^*$ in tensor linear regression.

4.1 Introduction

Many modern machine learning and data science problems involve high dimensional multiway (tensor) structures. Examples of problems wherein tensors have found applications include (but are not limited to) recommendation systems [74, 104–106], mixture and topic modeling [107, 108], deep learning [73, 109–113], multilinear subspace learning [114, 115], and speech source separation [116, 117]. As we discussed in Chapter 3, taking advantage of the structured and the higher order correlations in tensor structures reduces the dimensionality of the problem and can result in more accurate predictions or estimations (or equivalently lowering sample complexity required to obtain a target accuracy). In this chapter, we study low-rank tensor linear regression (TLR), a class of supervised learning models that aim to exploit the tensor structure in the predictors and the regression parameters to allow for solving high dimensional linear regression problems accurately when the number of observations is only a small fraction of the

number predictors. Tensor linear regression has application in many areas including multitask learning [88], complex network analysis [118], and neuroimaging data analysis [119, 120].

Tensor regression models methods share the assumption that the model parameters form a high order tensor and there exists a low dimensional factorization for the regression tensor. The model we consider here is based on CANDECOMP/PARAFAC (CP) decompostion of tensors which allows for explicit accounting of interdependencies along different modes of tensor arrays.

Here, we focus on providing lower bounds on minimax risk of estimating the regression tensor using any estimator. By comparing these bounds to those of standard (i.e. vectorized) linear regression, we show the benefits of exploiting the tensor structure in the linear regression problem with tensor-structured predictors and parameters. For standard linear regression, the lower bound on the minimax risk of estimating a parameter vector in $\mathbb{R}^m$ is $\Omega\big(\frac{m\sigma^2}{L\|\Sigma_x\|_2}\big)$ where Σ_x is the covariance matrix of the predictor vector and σ is the noise variance [121]. Therefore, by vectorizing tensor data samples in $\mathbb{R}^{m_1\times m_2\times\cdots\times m_N}$, we have a lower bound on the worst case MSE of estimating the true model parameters in form of

$$\epsilon^* \geq \Omega\Big(\frac{\prod_{n=1}^{N} m_n\sigma^2}{L\|\Sigma_x\|_2}\Big). \tag{4.1}$$

In contrast, we show that when the spatial structure of data is preserved and a CP-rank-p structure is imposed on the parameter tensor, the minimax lower bound is reduced to

$$\epsilon^* \geq \Omega\Big(\frac{\sigma^2 p\sum_{n=1}^{N} m_n}{NL\|\Sigma_x\|_2}\Big). \tag{4.2}$$

4.1.1 Relation to Prior Work

Tensor decompositions have received a lot of attention in the recent years as a tool to avoid overparameterization in tensor data models [71, 72, 122]. The resulting more compact models tend to be more efficient with regards to storage, computational complexity,

and generalization performance. This has motivated the use of tensor decompositions in a variety of areas, including deep learning [73, 109], collaborative filtering [74, 105], multilinear subspace learning [114], source separation [116], topic modeling [108], and many other works [74]. In the recent years, tensor decompositions have also received attention in regression problems such as neuroimaging data analysis [119, 120] where data is tensor structured and high dimensional but the sample size is relatively small.

While some works in the literature consider tensor linear regression problems with tensor responses [123–126], our focus in this work is on the model with scalar response. A majority of the works on tensor linear regression (TLR) focus on the algorithmic aspects of TLR and developing efficient solvers for the problem under different settings [119, 120, 127–134]. In contrast, fewer works study the theoretical aspect of the TLR problem in terms of the fundamental limits of the TLR models. Wimalawarne et al. study regularized tensor regression with different choices of the regularization term and derive excess risk bounds for each regularized model. Zhou et al. [119] study the conditions for local identifiablity of the true parameter tensor in CP-based TLR model. The CP-based model assumes that the parameter tensor has a low-CP-rank structure (i.e. the CP-rank is at most p for some small p). The authors show that the required number of samples for idetifiability is reduced from $\Omega\big(\prod_{n=1}^{N} m_n\big)$ to $\Omega\big(p\sum_{n=1}^{N} m_n\big)$. In the same vein, Li et al. [120] investigate the TLR model based on Tucker decomposition, where the assumption is that the Tucker-rank of the parameter tensor is small. The authors show that, similar to the CP-based model, the required sample complexity for local identifiability of the true tensor is a linear function of the number of parameters (degrees of freedom) in the model. In terms of works on the minimax risk in TLR, Suzuki [135] obtain results for CP-based tensor completion (which can be though of as a special case of tensor regression, where the elements of predictor tensors are all zero except for a single element with value 1). Their result, however, does not trivially extend to the general TLR problem. To the best of our knowledge, our work is the first result on minimax risk in the general CP-based TLR model.

Our approach for deriving minimax result is a well-established information theoretic method that was first proposed by Khas'minskii [136] and was later developed further

by other researchers [137, 138]. Specifically, because of our analysis of tensor-structured parameters, our proofs borrow many analytical tools from the works by Jung et al. [139] and Shakeri et al. [64, 140].

4.2 Preliminaries and Problem Statement

4.2.1 Notation and Definitions

We use underlined bold upper-case ($\underline{\mathbf{A}}$), bold upper-case ($\mathbf{A}$), bold lower-case ($\mathbf{a}$), and lower-case (a) letters to denote tensors, matrices, vectors, and scalars, respectively. For any integer p, we define $[p] \triangleq \{1, 2, \cdots, p\}$. We denote by $\{\mathbf{A}_n\}_{n=1}^N$ the set $\{\mathbf{A}_1, \cdots, \mathbf{A}_N\}$. We drop the range indicators if they are clear from the context.

Norms and inner products: We denote by $\|\mathbf{v}\|_p$ the ℓ_p norm of vector $\mathbf{v}$, while we use $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_F$ to denote the spectral and the Frobenius norm of matrix $\mathbf{A}$, respectively. We define by $\|\mathbf{A}\|_0$ the number of nonzero elements of matrix (or vector) $\mathbf{A}$. We define the inner product of two tensors (or matrices) $\underline{\mathbf{A}}$ and $\underline{\mathbf{B}}$ as $\langle \underline{\mathbf{A}}, \underline{\mathbf{B}} \rangle \triangleq \langle \text{vec}(\underline{\mathbf{A}}), \text{vec}(\underline{\mathbf{B}}) \rangle$ where $\text{vec}(\cdot)$ is the vectorization operator.

Matrix products: We denote the Hadamard product (element-wise product) of matrices $\mathbf{A} \in \mathbb{R}^{m \times p}$ and $\mathbf{B} \in \mathbb{R}^{m \times p}$ by $\mathbf{A} \bullet \mathbf{B} \in \mathbb{R}^{m \times p}$. We denote by $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{m_1 m_2 \times p_1 p_2}$ the Kronecker product of matrices $\mathbf{A} \in \mathbb{R}^{m_1 \times p_1}$ and $\mathbf{B} \in \mathbb{R}^{m_2 \times p_2}$. We use $\bigotimes_{n=1}^N \mathbf{A}_i \triangleq \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_N$ for the Kronecker product of N matrices. The Khatri-Rao product of two matrices $\mathbf{C} \in \mathbb{R}^{m_1 \times p}$ and $\mathbf{D} \in \mathbb{R}^{m_2 \times p}$ is denoted by $\mathbf{C} \odot \mathbf{D} \in \mathbb{R}^{m_1 m_2 \times p}$. We use $\bigodot_{n=1}^N \mathbf{C}_i \triangleq \mathbf{C}_1 \odot \mathbf{C}_2 \odot \cdots \odot \mathbf{C}_N$ for the Khatri-Rao product of N matrices.

Definitions for tensors: We denote the outer product (tensor product) of vectors by $\circ$, while $\times_n$ denotes the mode-n product between a tensor and a matrix. An N-way tensor is rank-1 if it can be written as outer product of N vectors: $\mathbf{v}_1 \circ \cdots \circ \mathbf{v}_N$. A *superdiagonal* tensor $\underline{\mathbf{S}} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_N}$ is a tensor with all zero elements except for the superdiagonal elements, i.e., the elements indexed by $(i_1, i_2, \cdots, i_N)$ such that $i_1 = i_2 = \cdots = i_N$. We denote by $\mathcal{S}_p^N$ the set of all N-way superdiagonal tensors in $\mathbb{R}^{p \times p \times \cdots \times p}$.

Throughout this chapter, by the rank of a tensor, rank($\underline{\mathbf{A}}$), we mean the CP-rank

of $\underline{\mathbf{A}}$, the minimum number of rank-1 tensors that construct $\underline{\mathbf{A}}$ as their sum. The *CP decomposition* (CPD), decomposes a tensor into sum of its rank-1 tensor components. The *Tucker decomposition* factorizes an N-way tensor $\underline{\mathbf{A}} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_N}$ as $\underline{\mathbf{A}} = \underline{\mathbf{X}} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times_3 \cdots \times_N \mathbf{D}_N$, where $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_N}$ denotes the core tensor and $\mathbf{D}_n \in \mathbb{R}^{m_n \times p_n}$ denote factor matrices along the n-th mode of $\underline{\mathbf{A}}$ for $n \in [N]$.

Notations for functions and spaces: We denote by $\mathcal{D}_{m \times p}$ the oblique manifold in $\mathbb{R}^{m \times p}$; the manifold of matrices with unit-norm columns: $\mathcal{D}_{m \times p} \triangleq \{\mathbf{D} \in \mathbb{R}^{m \times p} | \forall j \in [p],\ \mathbf{d}_j^T \mathbf{d}_j = 1\}$.

4.2.2 Low-Rank Tensor Linear Regression

In tensor linear regression we assume that each scalar response $y_l \in \mathbb{R}$ is generated according to

$$y_l = \langle \underline{\mathbf{B}}^*, \mathbf{X}_l \rangle + \epsilon_l, \tag{4.3}$$

where $\underline{\mathbf{B}}^* \in \mathbb{R}^{m_1 \times \cdots \times m_N}$ is the true underlying regression tensor, $\{\underline{\mathbf{X}}_l \in \mathbb{R}^{m_1 \times \cdots \times m_N}\}_{l=1}^L$ is the corresponding (randomly generated) predictor tensor, and $\epsilon_l \in \mathbb{R}$ is the observation noise. In order to explicitly account for the tensor structure in the coefficient tensor, we adopt a tensor-factorization based model. While there are many ways to factorize a tensor, we consider a model based on well-known tensor factorization called the CANDECOMP/PARAFAC (CP) decomposition (Also known as the tensor rank decomposition[1]). The CP decomposition factorizes a tensor into sum of its rank-1 tensor components:

$$\underline{\mathbf{A}} = \sum_{k=1}^{r} \lambda_k \mathbf{a}_1^k \circ \cdots \circ \mathbf{a}_N^k, \tag{4.4}$$

[1]Sometimes CP decomposition is is considered as the generalization of the tensor rank decomposition where the number of the terms can be larger than the rank of the tensor. In this Chapter, however, we use only consider the case with minimal number of terms (minimal CP decomposition) and use the two names interchangeably.

where $\mathbf{a}_n^k$ is a unit-norm vector for $k \in [r]$ and $n \in [N]$. When the number of terms r is minimal in the above expression, then r is called the CP-rank of the tensor[2]. Therefore we can write the underlying low-rank coefficient tensor $\underline{\mathbf{B}}^*$ as

$$\underline{\mathbf{B}}^* = \sum_{j=1}^{p} g_j^* \;\; \mathbf{b}_{1,j}^* \circ \cdots \circ \mathbf{b}_{N,j}^*,$$

where p is small ($p \ll \min_i \prod_{n \neq i} m_n$). We can write the CP decomposition of $\underline{\mathbf{B}}^*$ in the following more compressed way

$$\underline{\mathbf{B}}^* = \underline{\mathbf{G}}^* \times_1 \boldsymbol{\beta}_1^* \cdots \times_N \boldsymbol{\beta}_N^* \in \mathcal{B}, \tag{4.5}$$

where the set $\mathcal{B}$ is defined as

$$\mathcal{B} = \{\underline{\mathbf{G}} \times_1 \boldsymbol{\beta}_1 \times_2 \cdots \times_N \boldsymbol{\beta}_N | \underline{\mathbf{G}} \in \mathcal{S}_p^N, \; \boldsymbol{\beta}_n \in \mathcal{D}_{m_n \times p}, \; \forall n \in [N]\}, \tag{4.6}$$

where $\mathcal{S}_p^N$ is the set of N-way superdiagonal tensors in $\mathbb{R}^{p \times \cdots \times p}$ and $\mathcal{D}_{m_n \times p}$ is the oblique manifold[3] in $\mathbb{R}^{m_n \times p}$. We can then express the tensor regression model in the following way:

$$\begin{aligned} y_l &= \langle \mathrm{vec}(\underline{\mathbf{B}}^*), \mathrm{vec}(\mathbf{X}_l) \rangle + \epsilon_l \\ &= \langle (\boldsymbol{\beta}_N \otimes \cdots \otimes \boldsymbol{\beta}_1) \mathrm{vec}(\underline{\mathbf{G}}^*), \mathrm{vec}(\mathbf{X}_l) \rangle + \epsilon_l \\ &= \left\langle \bigotimes_{n=N}^{1} \boldsymbol{\beta}_n^* \;\; \mathrm{vec}(\underline{\mathbf{G}}^*), \mathrm{vec}(\mathbf{X}_l) \right\rangle + \epsilon_l. \end{aligned} \tag{4.7}$$

Therefore, the problem reduces from estimating $\underline{\mathbf{B}}^*$ to estimating $\{\boldsymbol{\beta}_n^*\}_{n=1}^N$ and the superdiagonal elements of $\underline{\mathbf{G}}^*$, considerably reducing the number of parameters to be estimated in the problem.

---

[2]Throughout this Chapter, by the rank of a tensor, rank($\underline{\mathbf{A}}$), we mean the CP-rank of $\underline{\mathbf{A}}$.

[3]The unit-norm condition on columns of $\{\boldsymbol{\beta}_n^{t_n}\}$ is to simplify the analysis by avoiding the ambiguity stemming from the invariance of CP decomposition to scaling of factor matrices.

4.2.3 Minimax Risk

Many learning problems, including linear regression, boil down to estimation problem where some model parameters need to be estimated. Minimax risk, defined as the lowest risk achievable by *any* estimator in the worst possible case allowed in an estimation problem, is an important theoretical tool in understanding the fundamental limits of such learning problems. These fundamental limits are in terms of bounds on the performance of estimation (or optimization) algorithms. Such bounds are important in understanding whether the existing algorithms to solve a problem are optimal (with respect to a certain metric) or one can still develop more efficient algorithms in that metric. Moreover, these bounds can be used to compare different approaches in modeling a learning problem and understanding the benefits of exploiting our knowledge of certain structures in data or the underlying generative model of the data.

Let us now formally define the minimax risk in an estimation problem. let $\mathcal{P}$ denote a family of distributions on a sample space $\mathcal{X}$, and let $\theta : \mathcal{P} \to \Theta$ denote a mapping $P \to \theta(P)$. The goal is to estimate the true model parameter $\theta(P)$ based on i.i.d. observations $X_1 \in \mathcal{X}$ drawn from an unknown distribution $P \in \mathcal{P}$. To measure the error of an estimator of parameter θ, we employ the (semi-)metric $\rho : \Theta \times \Theta \to \mathbb{R}_+$. Given a set of observations $X_1, \cdots, X_L$, the minimax risk achieved by any estimator of θ is defined as

$$\inf_{\hat{\theta} \in \Theta} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\rho\big(\hat{\theta}(X_1, \cdots, X_L), \theta(P)\big) \right], \tag{4.8}$$

where the supremum (representing worst case scenario) is taken over all possible distributions in $\mathcal{P}$ and the infimum is taken over all estimators.

Fano's inequality: A common technique for finding lower bounds on minimax risk in estimation problems, which we adopt in this work, involves connecting the estimation problem to a multiple hypothesis testing problem. Fano's inequality provides lower bounds on the error in a multiple hypothesis testing problem, is an essential component of this technique.

Lemma 12 (Fano's Inequality)**.** *Let V be a random variable taking values in a finite*

set $\mathcal{V}$ with cardinality $|\mathcal{V}| \geq 2$. Consider any Markov chain $V \to X \to \hat{V}$. Let e denote the occurrence of $\hat{V} \neq V$. Then, we have

$$H(e) + \mathbb{P}(e)\log(|\mathcal{V}| - 1) \geq H(V|\hat{V}), \tag{4.9}$$

where $H(A)$ is the entropy of random variable A and $H(A|B)$ is the conditional entropy of A given B.

4.3 Minimax Risk of Tensor Linear Regression

We wish to put a lower bound on the minimax risk of estimators for estimating the rank-p coefficients tensor $\underline{\mathbf{B}}^*$ in the low-rank tenor regression problem, based on observations $(\mathbf{X}_l, y_l)$. As mentioned earlier, here we consider the CP-based model where the observations are generated according to

$$y_l = \langle \bigotimes_{n\in[N]} \boldsymbol{\beta}_n^* \ \mathrm{vec}(\underline{\mathbf{G}}^*), \mathrm{vec}(\mathbf{X}_l)\rangle + \epsilon_l. \tag{4.10}$$

We further make the following assumptions on the generating model:

Assumption 5 (Model assumptions)**.** *We assume that in the generating model* (4.10)*, we have*

1. $\underline{\mathbf{B}}^* \in \mathbb{R}^{m_1\times\cdots\times m_N}$ *is the true, unknown N-way coefficient tensor.*
2. $\underline{\mathbf{X}}_l \in \mathbb{R}^{m_1\times\cdots\times m_N}$ *is the N-way covariate tensor (tensor predictor) with some known distribution and covariance Σ_x,*
3. $\epsilon_l \sim \mathcal{N}(0, \sigma^2)$ *is zero-mean Gaussian noise, independent and uncorrelated to the class parameter $\underline{\mathbf{B}}^*$ and predictor variable $\underline{\mathbf{X}}_l$.*

Under the CP-based model, the tensor linear regression problem reduces to estimating $\{\boldsymbol{\beta}_n^*\}_{n=1}^N$ and the superdiagonal core tensor $\underline{\mathbf{G}}^*$. For the ease of analysis, we assume that we have the *a priori knowledge of the superdiagonal elements of $\underline{\boldsymbol{G}}^*$*, and the N factor matrices remain to be estimated.

The analysis that we provide here is *local* in that we assume that the true coefficients tensor $\underline{\mathbf{B}}^*$ lies in a neighborhood of radius r around a fixed reference tensor

$$\underline{\mathbf{B}}^0 = \underline{\mathbf{G}}^0 \times_1 \boldsymbol{\beta}_1^0 \cdots \times_N \boldsymbol{\beta}_N^0 \in \mathcal{B}, \tag{4.11}$$

denoted by

$$\mathcal{B}_r = \{\underline{\mathbf{B}} \in \mathcal{B} | \rho(\underline{\mathbf{B}}, \underline{\mathbf{B}}^0) < r\}. \tag{4.12}$$

where $\mathcal{B}$ is defined in (4.6). In this work, we choose the semi-metric $\rho(\underline{\mathbf{B}}, \underline{\mathbf{B}}')$ is chosen to be $\|\underline{\mathbf{B}} - \underline{\mathbf{B}}'\|_F^2$. This local analysis avoids ambiguity issues intrinsic to tensor regression problem due to non-uniqueness of CP decomposition. It is trivial to show, however, that lower bounds on the minimax risk in the local setting also apply to the global setting ($r \to \infty$).

We define the minimax risk as the worst-case mean squared error (MSE) that can be obtained by the best rank-p tensor estimator $\hat{\underline{\mathbf{B}}}$:

$$\epsilon^* \triangleq \inf_{\hat{\underline{\mathbf{B}}}} \sup_{\underline{\mathbf{B}} \in \mathcal{B}_r} \mathbb{E}_{\underline{\mathbf{B}}} \left[\|\hat{\underline{\mathbf{B}}} - \underline{\mathbf{B}}\|_F^2 \right]. \tag{4.13}$$

Our goal here is to provide a lower bound on ϵ^* using an information-theoretic methodology that we describe in detail next.

4.3.1 Our Approach

We will follow the information-theoretic approach known as Fano's method [121, 136–138]. First, we reduce the problem of estimating $\underline{\mathbf{B}}^*$ to a multiple hypothesis testing problem between a finite family of coefficient tensors: $\mathbb{B}_T = \{\underline{\mathbf{B}}^1, ..., \underline{\mathbf{B}}^T\} \in \mathcal{B}_r$. In this approach, we assume that the true dictionary is chosen uniformly at random from the set $\mathbb{B}_T$. If there is an estimator with small enough worst case MSE, then we can use this estimator to solve the multiple hypothesis testing problem. We then can use Fano's inequality that lower bounds the error in multiple hypothesis testing problem which we will use to provide a lower bound on the worst case MSE of the best estimator, i.e.

the minimax error. Now the question becomes how to set up the multiple hypothesis testing problem such that we can obtain tight lower bounds on the minimax error in the estimation problem. We discuss this next.

In the hypothesis testing problem, we assume that nature chooses a t^* uniformly at random from the index set $[T]$. The task is now detecting the true coefficient tensor $\underline{\mathbf{B}}_{t^*} \in \mathbb{B}_T$ using observations $(\underline{\mathbf{X}}_l, y_l)$. The following lemma, which is an adaptation of Proposition 2.3 in Duchi [121], formalizes the relation between this hypothesis testing problem and the original tensor regression problem.

Lemma 13. *Consider the regression model* (4.10) *with minimax risk* ϵ^* *defined in* (4.13). *Let* $\hat{\underline{\mathbf{B}}}$ *denote*[4] *an arbitrary estimator for the true coefficient tensor* $\underline{\mathbf{B}}^*$ *defined in model* (4.10). *Moreover, consider the set* $\mathbb{B}_T = \{\underline{\mathbf{B}}_1, \cdots, \underline{\mathbf{B}}_T\} \subset \mathcal{B}_r$. *consider the multiple hypothesis testing problem where the the true coefficient tensor is chosen uniformly at random from* $\mathbb{B}_T$ *and is indexed by* t^*. *Let* $\hat{t}(\hat{\underline{\mathbf{B}}})$ *denote a minimum distance detector such that*

$$\hat{t}(\hat{\underline{\mathbf{B}}}) = \underset{t \in [T]}{\operatorname{argmin}} \|\hat{\underline{\mathbf{B}}} - \underline{\mathbf{B}}_t\|_F^2,$$

where $\mathbf{B}_t \in \mathcal{B}_r$ *for all* $t \in [T]$. *Then, we have*

$$\epsilon^* \geq \min_{t,t' \in [T]} \|\underline{\mathbf{B}}_t - \underline{\mathbf{B}}_{t'}\|_F^2 \cdot \inf_{\hat{\underline{\mathbf{B}}}} \mathbb{P}(\hat{t}(\hat{\underline{\mathbf{B}}}) \neq t^*). \tag{4.14}$$

Lemma 13 indicates that in order to obtain tight lower bounds on the minimax risk, we need to construct $\mathbb{B}_T$ such that the distance between any two tensors in $\mathbb{B}_T$ is large (maximizing the first term in the lower bound (4.14)) while the hypothesis testing problem is also hard, i.e., two distinct coefficient tensors produce similar observations (maximizing the second term in the lower bound (4.14)).

More details of the construction will be provided in the proof of Theorem 8, our main result.

---

[4]Throughout this chapter we suppress the dependence $\hat{\underline{\mathbf{B}}}$ on the random observations $\{(\underline{\mathbf{X}}, \mathbf{y})\}$.

4.3.2 Main Result

Here, we first state the main result of this chapter on the minimax risk of CP-based tensor linear regression in Theorem 8 and then discuss the proof of the theorem in both big picture and detail.

Theorem 8. *Consider a tensor linear regression problem with L i.i.d. observations generated according to model* (4.10) *where the core tensor $\underline{\mathbf{G}}^*$ is known. Fix a reference tensor $\underline{\mathbf{B}}^0$ satisfying* (4.11) *and a positive constant r and suppose that the true parameter tensor $\underline{\mathbf{B}}^*$ in model* (4.10) *belongs to $\mathcal{B}_r$ defined in* (4.12)*. Further, assume that Assumption 5 holds. Then, the minimax lower bound ϵ^* defined in* (4.13) *can be bounded as follows*

$$\epsilon^* \geq \frac{t}{4} \min \Big\{ \frac{\|\underline{\mathbf{G}}^*\|_F^2}{\kappa^2}, \ \frac{\|\underline{\mathbf{G}}^*\|_F^2 r^2}{2pN\kappa^2}, \ \frac{\Big(c_1 \ p\sum_{n=1}^{N}(m_n-1) + N\left(1-\frac{1}{2}\log_2 N\right) - 2\Big)\sigma^2}{4NL\|\Sigma_x\|_2 p} \Big\}, \tag{4.15}$$

where $t \leq \min_{n\in[N]} \frac{2}{(m_n-1)p}$ and $0 < c_1 < 1$ and $\kappa > 1$.

We first provide an outline of the proof, then provide the formal proof.

Outline of the proof of Theorem 8. We set up a multiway hypothesis testing problem by constructing a set of T distinct tensors in the parameter space $\mathbb{B}_T = \{\underline{\mathbf{B}}^1, \cdots, \underline{\mathbf{B}}^T\} \in \mathcal{B}_r$ where

$$\mathcal{B}_r \triangleq \{\underline{\mathbf{B}} \in \mathcal{B} |\ \|\underline{\mathbf{B}} - \underline{\mathbf{B}}^0\|_F < r\}, \tag{4.16}$$

such that

$$\min_{t,t'\in[T]} \|\underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'}\|_F^2 \geq 2\delta. \tag{4.17}$$

for some positive value δ. The true coefficient tensor, indexed by t^*, then is chosen uniformly at random from the set $\mathbb{B}_T$. It follows from the generating model (4.7) that

the responses y generated using this parameter tensor follow a Gaussian distribution, conditioned on the predictor tensor $\underline{\mathbf{X}}$ whose distribution is known. This allows us to provide an upper bound on the mutual information $I(t^*; \mathbf{y}|\underline{\mathbf{X}})$ in terms of $\{m_n\}_{n=1}^N$, p, N, σ, Σ_x, r, and some parameter $\epsilon > 0$ that we will connect to ϵ^*. Since we also have the lower bound

$$I(t^*; \mathbf{y}|\underline{\mathbf{X}}) \geq (1 - \mathbb{P}(\hat{t} \neq t^*)) \log_2 T - 1 \tag{4.18}$$

from Fano's inequality and data processing inequality, we will obtain a relation in the following form that will allow us to provide lower bound on the minimax error:

$$(1 - \mathbb{P}(\hat{t} \neq t^*)) \log_2 T - 1 \leq \mathbb{I}(t^*; \mathbf{y}|\underline{\mathbf{X}}) \leq h(\epsilon^*), \tag{4.19}$$

where $h(\cdot)$ is a linear function. We choose δ to be the smallest value large enough that $\mathbb{P}_e$ is less then an arbitrary constant, making the lower bound on $I(t^*; \mathbf{y}|\underline{\mathbf{X}})$ only a function of T. By choosing the value of δ not larger than required, we ensure that the maximum distance between any two points, and therefore maximum KL divergence between any two distributions is small (i.e. upper bound on $I(t^*; \mathbf{y}|\underline{\mathbf{X}})$ is tight). Finally, we can compare the lower bound and the upper bound on $I(t^*; y|\mathbf{X})$ to get the desired result. □

Next, we present the formal proof of Theorem 8. In order to prove the minimax lower bound in Theorem 8, we employ the results of Lemmas 14- 16. Proofs of Lemmas 14 and 15 are provided at the end of this chapter, in Section 4.5. Proof of Lemma 16 is a simple adaptation of that of Lemma IV.4 in Jung et al. [139].

Proof of Theorem 8. The proof of Theorem 8 is based reducing the estimation problem under consideration to a multiple hypothesis testing problem. The hypothesis test is performed among the members of $\mathbb{B}_T = \{\underline{\mathbf{B}}^1, \underline{\mathbf{B}}^2, \cdots, \underline{\mathbf{B}}^T\} \in \mathcal{B}_r$ where $\mathcal{B}_r$ is defined in (4.12). We assume that the true dictionary, indexed by t^*, is chosen uniformly at random from the set $\mathbb{B}_T$.

It follows from Lemma 13 that in order to obtain a tight lower bound on minimax

error ϵ^*, we need to construct $\mathbb{B}_T$ such that the minimum pairwise distance of its elements is large, i.e.,

$$\min_{t,t' \in [T]} \|\underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'}\|_F^2 \geq 2\delta, \tag{4.20}$$

for some large $\delta > 0$, while the KL divergence between pairs of conditional distributions of the response variables, denoted by $D_{KL}\big(f_{\underline{\mathbf{B}}_t}(\mathbf{y}|\underline{\mathbf{X}})||f_{\underline{\mathbf{B}}_{t'}}(\mathbf{y}|\underline{\mathbf{X}})\big)$, is small, i.e.

$$D_{KL}\big(f_{\underline{\mathbf{B}}_t}(\mathbf{y}|\underline{\mathbf{X}})||f_{\underline{\mathbf{B}}_{t'}}(\mathbf{y}|\underline{\mathbf{X}})\big) < \eta \tag{4.21}$$

for some small $\eta > 0$. To find sufficient condition on cardinality T such that a construction satisfying conditions (4.20) and (4.21) exists, we rely on the following lemma.

Lemma 14. *Consider a constant $\alpha \geq 2$, an integer p, and N positive integers m_n for $n \in [N]$. Consider N positive integers T_n for $n \in [N]$ such that*

$$\log_2(T_n) < \frac{m_n p (2-\alpha)^2}{4\alpha^2 \log(2)} - \frac{1}{2}\log_2(N) + 1 \tag{4.22}$$

for all $n \in [N]$. Then, there exist N sets in form of $\mathcal{A}_n = \{\mathbf{A}_t^n \in \mathbb{R}^{m_n \times p} : t \in [T_n]\}$ for $n \in [N]$ where each set is comprised of binary matrices

$$\mathbf{A}_n^t \in \left\{-\frac{1}{\sqrt{m_n}}, \frac{1}{\sqrt{m_n}}\right\}^{m_n \times p} \tag{4.23}$$

satisfying

$$\left\|\mathbf{A}_n^t - \mathbf{A}_n^{t'}\right\|_0 \geq \frac{m_n p}{\alpha} \tag{4.24}$$

for all $t, t' \in T_n$.

Next, we derive sufficient conditions on cardinality T and parameter ϵ of the construction such that we guarantee existence of the construction satisfying (4.20) and (4.21). We also specify the values of δ and η in the lower bound (4.20) and the upper bound (4.21) in terms of the parameters of our construction.

Lemma 15. *Consider the tensor regression generative model in* (4.10). *Fix* $r > 0$ *and a reference tensor according to* (4.11). *Then there exists a collection of* L *tensors* $\mathbb{B}_T = \{\underline{\mathbf{B}}_1, \underline{\mathbf{B}}_2, \ldots, \underline{\mathbf{B}}_T\} \subset \mathcal{B}_r$ *of cardinality* $T = 2^{\sum_{n\in[N]} \frac{(m_n-1)p(2-\alpha)^2}{4\alpha^2\log(2)} - \frac{N}{2}\log_2(N)+N}$, *such that for any* $0 < t \leq \min_{n\in[N]} \frac{2}{(m_n-1)p}$ *and* $\varepsilon > 0$ *satisfying*

$$\varepsilon < \min\{1, \frac{r^2}{2pN}\}, \tag{4.25}$$

we have

$$\left\|\underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'}\right\|_F^2 \geq \frac{2t\varepsilon}{\kappa^2}\|\underline{\mathbf{G}}^*\|_F^2 \tag{4.26}$$

for all pairs of $t, t' \in [T]$, $t \neq t'$, *and*

$$\mathbb{I}(t^*; \mathbf{y}|\mathbf{X}) \leq \frac{2LNp\|\underline{\mathbf{G}}^*\|_F^2 \left\|\Sigma_x\right\|_2}{\sigma^2}\varepsilon. \tag{4.27}$$

As we discussed in the outline of the proof of Theorem 8, our approach is based on connecting the minimax error to model parameters by providing an upper bound and a lower bound on the conditional Mutual information $\mathbb{I}(t^*; \mathbf{y}|\mathbf{X})$. While in Lemma 15 we obtain an upper bound on the $\mathbb{I}(t^*; \mathbf{y}|\mathbf{X})$, we do not explicitly obtain a lower bound on this quantity. Instead, Lemma 15 gives a lower bound on the distance between any two points in our construction. In the following lemma, however, we connect lower bounds on pairwise distances in the construction to a lower bound on the conditional Mutual information.

Lemma 16 (Lower bound on MI)**.** *Consider the linear regression model in* (4.10) *and suppose that the minimax risk* $\varepsilon^* \leq \frac{\delta}{4}$ *for some* $\delta > 0$. *Assume that there exists a finite set of* L *distinct coefficient tensors* $\mathbb{B}_L = \{\boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_L\} \subset \mathcal{N}_r(\boldsymbol{\beta}_0)$ *such that* $\min_{t,t'\in[T]} \|\underline{\mathbf{B}}_t - \underline{\mathbf{B}}_{t'}\|_F^2 \geq 2\delta$. *Then, we have*

$$\frac{1}{2}\log_2 T - 1 \leq \mathbb{I}(t^*; \mathbf{y}|\underline{\mathbf{X}}). \tag{4.28}$$

It follows from Lemma 15 that for any $\varepsilon > 0$ satisfying condition (4.25), there exists

an set of $\mathbb{B}_T \subset \mathcal{B}_r$ with cardinality $T = 2^{c_1 \sum_{n\in[N]}(m_n-1)p - \frac{N}{2}\log_2(N)+N}$ that satisfies (4.27), where $c_1 = \frac{(2-\alpha)^2}{4\alpha^2 \log(2)} < 1$ for some $\alpha > 2$. Moreover, Lemma 15 implies that if there exists an estimator with worst case MSE smaller than

$$\frac{t\|\underline{\mathbf{G}}^*\|_F^2}{4\kappa^2} \min\{1, \frac{r^2}{2pN}\},$$

then we can set ε such that $\frac{2t\|\underline{\mathbf{G}}\|_F^2}{\kappa^2}\varepsilon = 8\epsilon^*$. This means that there exists a $\delta > 0$ such that $\delta \geq 4\epsilon^*$ and $\min_{t,t'\in[T]} \|\underline{\mathbf{B}}_t - \underline{\mathbf{B}}_{t'}\|_F^2 \geq 2\delta$, which means that lower bound (4.28) also holds. Therefore, under these conditions we have

$$\frac{1}{2}\log_2 T - 1 \leq \mathbb{I}(t^*; \mathbf{y}|\mathbf{X}) \leq \frac{8\kappa^2\|\underline{\mathbf{G}}^*\|_F^2 LNp \left\|\Sigma_x\right\|_2}{t\|\underline{\mathbf{G}}^*\|_F^2\sigma^2}\epsilon^*,$$

or

$$c_1 \sum_{n\in[N]}(m_n-1)p - \frac{N}{2}\log_2(N) + N - 2 \leq \mathbb{I}(t^*; \mathbf{y}|\mathbf{X}) \leq \frac{16\kappa^2 LNp \left\|\Sigma_x\right\|_2}{t\sigma^2}\epsilon^*. \quad (4.29)$$

which gives us

$$\epsilon^* \geq \frac{t\sigma^2(c_1 \sum_{n\in[N]}(m_n-1)p - \frac{N}{2}\log_2(N) + N - 2)}{16\kappa^2 LNp \left\|\Sigma_x\right\|_2}, \quad (4.30)$$

which concludes the proof. □

4.3.3 Discussion

First we note that while our analysis is a local one, meaning that we only consider a neighborhood of radius r around a reference dictionary, our lower bound trivially holds for the global case where $r \to \infty$. Moreover, when sufficient number of samples are given, the minimax bound we provide has no dependence on the neighborhood size r, suggesting that the local nature of our analysis is not limiting.

Let us know investigate the lower bound (4.15) in Theorem 8 on the minimax risk in the tensor linear regression problem. Our bound, for sufficient large number of

samples, depends on the number of tensor order N, the parameters $(p\sum_{n=1}^{N} m_n)$, number of samples L, noise variance σ^2, and the covariance matrix of the predictors Σ_x. When we compare our minimax lower bound for CP-based TLR to the $\Omega\big(\frac{\sigma^2\prod_{n=1}^{N} m_n}{L\|\Sigma_x\|_2}\big)$ minimax lower bound of the ordinary (vectorized) linear regression, it becomes obvious that the dependence of the minimax error on the dimensions is reduced from $\Omega(\prod_{n=1}^{N} m_n)$ to $\Omega(p\sum_{n=1}^{N} m_n)$ when the tensor structure is taken into account in the tensor linear regression model. This confirms our intuition regarding benefits of exploiting tensor structure in linear regression problems with tensor. Specifically, by exploiting the tensor structure, it is possible to design estimators with improved worst case accuracy. Equivalently, we can present the improvement in terms of the sample complexity required to achieve a target expected worst case error: we show a reduction from $L \geq \Omega\big(\frac{\sigma^2\prod_{n=1}^{N} m_n}{\|\Sigma_x\|_2\epsilon}\big)$ to $L \geq \Omega\big(\frac{\sigma^2 p\sum_{n=1}^{N} m_n}{N\|\Sigma_x\|_2\epsilon}\big)$. This is especially important since in many applications of linear regression with tensor data, number of sample is quite small compared to the dimensions of the problem [12, 119]. Moreover, our bound shows inverse relation between minimax error and the sample size the SNR[5], which is desirable. We also see an inverse relation between ϵ^* and N for a fixed number of parameters.

4.4 Conclusion and Future Work

In this chapter we demonstrated the benefits of exploiting the tensor structure in linear regression problems with tensor data by quantifying the reduction in the minimax risk of estimating the true model parameters. We adopted a well-established information-theoretic approach to provide a lower bound on the minimax risk of estimating the true parameter tensor which we assumed it has a low CP rank. To this end, we reduced the estimation problem to a multiple hypothesis testing problem by constructing a finite set of low CP-rank tensors in a local neighborhood of a fixed reference dictionary and assuming the true tensor is chosen uniformly at random from this finite set. We then used Fano's inequality and properties of Gaussian distributions to provide upper and lower bounds on the mutual information between the observations and the parameter

[5]Note that we have $\text{SNR} = \frac{\text{Tr}(\Sigma_x)}{\sigma^2} \geq \frac{\|\Sigma_x\|_2}{\sigma^2}$

tensor in the model, which allowed us to find a lower bound on the minimax risk in the low-CP-rank tensor regression problem. To the best of our knowledge, this is the first result on lower bounds on minimax risk of estimating the tensor parameter in tensor linear regression problem.

In terms of future work, an obvious generalization is obtaining a minimax lower bound for CP-based model without a priori knowledge of the core tensor. Moreover, in this work we framed the CP model as a special case of the Tucker-based model. In the Tucker-based model the core tensor in the Tucker decomposition of the parameter tensor $\underline{\mathbf{B}}$ is not necessarily diagonal, and the dimensions of the core tensor can be different from one another. Providing minimax lower bounds for this more general case is a natural next step.

4.5 Proofs

In this section, we provide the proofs for Lemmas 14 and 15. To improve readability, the lemma statements are repeated here.

Lemma (Lemma 14)**.** *Consider a constant $\alpha \geq 2$, an integer p, and N positive integers m_n for $n \in [N]$. Consider N positive integers T_n for $n \in [N]$ such that*

$$\log_2(T_n) < \frac{m_n p (2-\alpha)^2}{4\alpha^2 \log(2)} - \frac{1}{2}\log_2(N) + 1$$

for all $n \in [N]$. Then, there exist N sets in form of $\varphi_n = \{\mathbf{\Phi}_n^{t_n} \in \mathbb{R}^{m_n \times p} : t_n \in [T_n]\}$ for $n \in [N]$ where each set is comprised of binary matrices

$$\mathbf{\Phi}_n^{t_n} \in \left\{ \frac{-1}{\sqrt{m_n}}, \frac{1}{\sqrt{m_n}} \right\}^{m_n \times p}$$

satisfying

$$\left\| \mathbf{\Phi}_n^{t_n} - \mathbf{\Phi}_n^{t'_n} \right\|_0 \geq \frac{m_n p}{\alpha},$$

for all $t_n, t'_n \in T_n$.

Proof of Lemma 14. consider N sets in form of $\varphi_n = \{\mathbf{\Phi}_n^{t_n} \in \mathbb{R}^{m_n \times p} : t_n \in [T_n]\}$ for $n \in [N]$. Let each set φ_n be a set of T_n matrices where each contains $m \times p$ independent and identically distributed random variables taking values $\pm\frac{1}{\sqrt{m_n}}$ uniformly. We $\bar{\mathbf{\Phi}}_n^{t_n,t'_n} \triangleq \mathbf{\Phi}_n^{t_n} \bullet \mathbf{\Phi}_n^{t'_n}$ be the Hadamard multiplication between $\mathbf{\Phi}_n^{t_n}$ and $\mathbf{\Phi}_n^{t'_n}$. Moreover, let $\bar{\phi}_{n,i}^{\ t_n,t'_n}$ be the i-th element of $\text{vec}(\bar{\mathbf{\Phi}}_n^{t_n,t'_n})$. We have

$$\left\|\mathbf{\Phi}_n^{t_n} - \mathbf{\Phi}_n^{t'_n}\right\|_0 = \frac{(m_n-1)}{2}\left[p - \sum_{i=1}^{m_n p} \bar{\phi}_{n,i}^{\ t_n,t'_n}\right]. \tag{4.31}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left(\left\|\mathbf{\Phi}_n^{t_n} - \mathbf{\Phi}_n^{t'_n}\right\|_0 \le \frac{(m_n-1)p}{\alpha}\right) &= \mathbb{P}\left(p - \sum_{i=1}^{(m_n-1)p} \bar{\phi}_{n,i}^{\ t_n,t'_n} \le \frac{2p}{\alpha}\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{(m_n-1)p} \bar{\phi}_{n,i}^{\ t_n,t'_n} \ge \frac{-p(2-\alpha)}{\alpha}\right) \\
&\overset{(a)}{\le} \exp\left[\frac{-2\left(\frac{p(2-\alpha)}{\alpha}\right)^2}{\sum_{i=1}^{(m_n-1)p}(\frac{2}{\sqrt{(m_n-1)}})^2}\right] \\
&= \exp\left[-\frac{(m_n-1)p(2-\alpha)^2}{2\alpha^2}\right],
\end{aligned} \tag{4.32}$$

where (a) follows from Hoeffding's inequality [45, 141] which we are allowed to use due to assumption $\alpha \ge 2$. Taking a union bound over all pairs $t_n, t'_n \in [T_n]$ for all $n \in N$:

$$\begin{aligned}
&\mathbb{P}\left(\exists n, t_n, t'_n : \left\|\mathbf{\Phi}_n^{t_n} - \mathbf{\Phi}_n^{t'_n}\right\|_0 \le \frac{(m_n-1)p}{\alpha}\right) \\
&\quad\le \sum_{n=1}^{N}\binom{T_n}{2}\exp\left[-\frac{(m_n-1)p(2-\alpha)^2}{2\alpha^2}\right] \\
&\quad\le N\max_{n\in N}\left(\frac{T_n^2}{2}\exp\left[-\frac{(m_n-1)p(2-\alpha)^2}{2\alpha^2}\right]\right) \\
&\quad= \max_{n\in N}\left(\exp\left[-\frac{(m_n-1)p(2-\alpha)^2}{2\alpha^2} + 2\log(T_n/2) + \log(N)\right]\right).
\end{aligned} \tag{4.33}$$

In order for the statement of the lemma to hold, we need the probability in (4.33) (the probability that the distance condition is violated for at least one pair of factor matrices

$\mathbf{\Phi}_n^{t_n}$ and $\mathbf{\Phi}_n^{t'_n}$) to be less than 1. That is,

$$-\frac{(m_n-1)p(2-\alpha)^2}{2\alpha^2}+2\log(T_n/2)+\log(N)<0, \quad \forall n\in N. \tag{4.34}$$

Therefore, we have

$$\log(T_n)<\frac{(m_n-1)p(2-\alpha)^2}{4\alpha^2}-\frac{\log(N)}{2}+\log(2), \quad \forall n\in N,$$

or

$$\log_2(T_n)<\frac{(m_n-1)p(2-\alpha)^2}{4\alpha^2\log(2)}-\frac{1}{2}\log_2(N)+1 \quad \forall n\in N. \tag{4.35}$$

We can further write this condition as

$$0<T_n<2^{\frac{(m_n-1)p(2-\alpha)^2}{4\alpha^2\log(2)}-\frac{1}{2}\log_2(N)+1}, \quad \forall n\in N. \tag{4.36}$$

This concludes the proof. □

Lemma (Lemma15)**.** *Consider the tensor regression generative model in* (4.10)*. Fix* $r>0$ *and a reference tensor according to* (4.11)*. Then there exists a collection of* L *tensors* $\mathbb{B}_T=\{\underline{\mathbf{B}}_1,\underline{\mathbf{B}}_2,\ldots,\underline{\mathbf{B}}_T\}\subset\mathcal{B}_r$ *of cardinality* $T=2^{\sum_{n\in[N]}\frac{(m_n-1)p(2-\alpha)^2}{4\alpha^2\log(2)}-\frac{N}{2}\log_2(N)+N}$*, such that for any* $0<t\leq\min_{n\in[N]}\frac{2}{(m_n-1)p}$ *and* $\varepsilon>0$ *satisfying*

$$\varepsilon<\min\{1,\frac{r^2}{2pN}\},$$

we have

$$\left\|\underline{\mathbf{B}}^t-\underline{\mathbf{B}}^{t'}\right\|_F^2\geq\frac{2t\varepsilon}{\kappa^2}\|\underline{\mathbf{G}}^*\|_F^2,$$

for all pairs of $t,t'\in[T]$*,* $t\neq t'$*, and*

$$\mathbb{I}(t^*;\mathbf{y}|\mathbf{X})\leq\frac{2LNp\|\underline{\mathbf{G}}^*\|_F^2\,\|\Sigma_x\|_2}{\sigma^2}\varepsilon.$$

Proof of Lemma 15. Let $\underline{\mathbf{B}}^0 = \underline{\mathbf{G}}^* \times_1 \boldsymbol{\beta}_1^0 \cdots \times_N \boldsymbol{\beta}_N^0$ be a rank-p reference tensor[6] such that the columns of $\boldsymbol{\beta}_n^0$ have unit norm for all $n \in [N]$. Let $\{\mathbf{U}_{n,j} \in \mathbb{R}^{m_n \times m_n}\}_{j=1}^p$ for $n \in [N]$ be arbitrary real unitary matrices such that

$$\mathbf{b}_{n,j}^0 = \mathbf{U}_{n,j} e_1, \ \forall n \in [N], \tag{4.37}$$

is the j-th column of $\boldsymbol{\beta}_n^0$. It follows from Lemma 14 that there exist N sets $\bar{\mathbb{B}}_n \subset \mathbb{R}^{(m_n-1)\times p}$ for $n \in [N]$ with elements

$$\bar{\boldsymbol{\beta}}_n^{t_n} \in \left\{ -\frac{1}{\sqrt{(m_n-1)}}, \frac{1}{\sqrt{(m_n-1)}} \right\}^{(m_n-1)\times p}, \ t_n \in [T_n],$$

such that $\min_{\bar{\boldsymbol{\beta}}_n^{t_n}, \bar{\boldsymbol{\beta}}_n^{t'_n} \in \bar{\mathbb{B}}_n} \left\| \bar{\boldsymbol{\beta}}_n^{t_n} - \bar{\boldsymbol{\beta}}_n^{t'_n} \right\|_0 \geq \frac{(m_n-1)p}{\alpha}$ for some $\alpha > 2$, if

$$\log_2(T_n) < \frac{(m_n-1)p(2-\alpha)^2}{4\alpha^2 \log(2)} - \frac{1}{2}\log_2(N) + 1,$$

for all $n \in N$. Now, we construct N sets $\tilde{\mathbb{B}}_n \subset \mathbb{R}^{m_n \times p}$ based on (and with the same cardinality as) the sets $\mathbb{B}^n \subset \mathbb{R}^{(m_n-1)\times p}$ in the following manner.

We construct each matrix $\tilde{\boldsymbol{\beta}}_n^{t_n} \in \tilde{\mathbb{B}}_n$ based on the matrix $\bar{\boldsymbol{\beta}}_n^{t_n} \in \bar{\mathbb{B}}_n$ and unitary matrices $\{\mathbf{U}_{n,j}\}_{j=1}^p$ such that the j-th column of $\tilde{\boldsymbol{\beta}}_n^{t_n}$ is given by

$$\tilde{\mathbf{b}}_{n,j}^{t_n} = \mathbf{U}_{n,j} \begin{bmatrix} 0 \\ \bar{\mathbf{b}}_{n,j}^{t_n} \end{bmatrix}, \ \forall n \in [N]. \tag{4.38}$$

Due to the constructions (4.38) and (4.37) and the fact that $\left\| \bar{\mathbf{b}}_{n,j}^{t_n} \right\|_2^2 = 1$, we have $\tilde{\mathbf{b}}_{n,j}^{t_n} \perp \mathbf{b}_{n,j}^0$, and $\left\| \mathbf{b}_{n,j}^0 \right\|_2^2 = 1$ and $\left\| \tilde{\mathbf{b}}_{n,j}^{t_n} \right\|_2^2 = 1$.

Now, we are ready to construct $\mathbb{B}_T$ with cardinality T where each $\underline{\mathbf{B}}_t \in \mathbb{B}_T$ is in form of

$$\underline{\mathbf{B}}^t = \underline{\mathbf{G}}^* \times_1 \boldsymbol{\beta}_1^{t_1} \cdots \times_N \boldsymbol{\beta}_N^{t_N}, \tag{4.39}$$

[6] remember that we assume knowledge of core tensor $\underline{\mathbf{G}}^*$

for $t \in T$ and $t_n \in T_n,\ n \in [N]$. We construct $\mathbb{B}_T$ such that

$$\boldsymbol{\beta}_n^{t_n} = \sqrt{1-\varepsilon}\boldsymbol{\beta}_n^0 + \sqrt{\varepsilon}\tilde{\boldsymbol{\beta}}_n^{t_n}, \tag{4.40}$$

for some $0 < \varepsilon < 1$. We will next derive conditions on ϵ and T. Throughout our analysis we utilize the fact that $\|\mathbf{b}_{n,j}^{t_n}\|_2 = 1$ and $\|\boldsymbol{\beta}_n^{t_n}\|_F = \sqrt{p}$.

Condition on T: We also derive the following condition on $T = |\mathbb{B}_T|$ based on condition (4.22) in the statement of Lemma 14 that

$$\begin{aligned} T = \prod_{n\in[N]} L_n &< \prod_{n\in[N]} 2^{\frac{(m_n-1)p(2-\alpha)^2}{4\alpha^2\log(2)} - \frac{1}{2}\log_2(N)+1} \\ &= 2^{\sum\limits_{n\in[N]} \frac{(m_n-1)p(2-\alpha)^2}{4\alpha^2\log(2)} - \frac{n}{2}\log_2(N)+N}, \end{aligned} \tag{4.41}$$

for some $\alpha > 2$.

Condition on ϵ: Next, in order to ensure that $\mathbb{B}_T \subseteq \mathcal{B}_r$, we show that $\|\underline{\mathbf{B}}^t - \underline{\mathbf{B}}^0\|_F^2$ for all $\underline{\mathbf{B}}^t \in \mathbb{B}_T$. We consider the following expansion of $\underline{\mathbf{B}}_t$ to aid in our future analysis:

$$\underline{\mathbf{B}}^t = \left[\sum_{\mathbf{i}\in\{0,1\}^N} a^{N-\|\mathbf{i}\|_1} b^{\|\mathbf{i}\|_1} \left(\bigotimes_{n\in[N]} \boldsymbol{F}_n^{i_n}\right)\right] \mathrm{vec}(\underline{\mathbf{G}}^*), \tag{4.42}$$

where $\mathbf{i} \triangleq (i_1, \ldots, i_N)$, $a \triangleq= \sqrt{1-\varepsilon}$, $b = \sqrt{\varepsilon}$, $\mathbf{F}_n^0 = \boldsymbol{\beta}_n^0$, and $\mathbf{F}_n^1 = \tilde{\boldsymbol{\beta}}_n^{t_n}$. We can proceed

with the following analysis:

$$
\begin{aligned}
&\left\|\underline{\mathbf{B}}^t - \underline{\mathbf{B}}^0\right\|_F^2 \\
&\quad= \left\| \left(\bigotimes_{n\in[N]} \boldsymbol{\beta}_n^0 \right) \operatorname{vec}(\underline{\mathbf{G}}^*) - \left[\sum_{\mathbf{i}\in\{0,1\}^N} a^{N-\|\mathbf{i}\|_1} b^{\|\mathbf{i}\|_1} \left(\bigotimes_{n\in[N]} \boldsymbol{F}_n^{i_n} \right) \right] \operatorname{vec}(\underline{\mathbf{G}}^*) \right\|_F^2 \\
&\quad\overset{(a)}{=} \left\| \left(\bigodot_{n\in[N]} \boldsymbol{\beta}_n^0 \right) \operatorname{vec}(\underline{\mathbf{G}}^*) - \left[\sum_{\mathbf{i}\in\{0,1\}^N} a^{N-\|\mathbf{i}\|_1} b^{\|\mathbf{i}\|_1} \left(\bigodot_{n\in[N]} \boldsymbol{F}_n^{i_n} \right) \right] \operatorname{vec}(\underline{\mathbf{G}}^*) \right\|_F^2 \\
&\quad= \left\| \left((1-a^N) \left(\bigodot_{n\in[N]} \boldsymbol{\beta}_n^0 \right) - \left(\bigodot_{n\in[N]} \boldsymbol{\beta}_n^0 \right) \right) \operatorname{vec}(\underline{\mathbf{G}}^*) \right\|_F^2 \\
&\quad\le \left\| (1-a^N) \left(\bigodot_{n\in[N]} \boldsymbol{\beta}_n^0 \right) - \left[\sum_{\substack{\mathbf{i}\in\{0,1\}^N \\ \|\mathbf{i}\|_1\neq 0}} a^{N-\|\mathbf{i}\|_1} b^{\|\mathbf{i}\|_1} \left(\bigodot_{n\in[N]} \boldsymbol{F}_n^{i_n} \right) \right] \right\|_F^2 \|\operatorname{vec}(\underline{\mathbf{G}}^*)\|_2^2 \\
&\quad= \left[(1-a^N)^2 \left\| \left(\bigodot_{n\in[N]} \boldsymbol{\beta}_n^0 \right) \right\|_F^2 + \sum_{\substack{i\in\{0,1\}^N \\ \|i\|_1\neq 0}} a^{2(N-\|\mathbf{i}\|_1)} b^{2\|\mathbf{i}\|_1} \left\| \bigodot_{n\in[N]} \boldsymbol{F}_n^{i_n} \right\|_F^2 \right] \|\underline{\mathbf{G}}^*\|_F^2 ,
\end{aligned}
$$

where (a) follows from the fact that $\underline{\mathbf{G}}^*$ is superdiagonal. For the first term in the bracket, i.e. $(1-a^N)^2 \left\| \left(\bigodot_{n\in[N]} \boldsymbol{\beta}_n^0 \right) \right\|_F^2$ we have

$$
\begin{aligned}
(1-a^N)^2 \left\| \left(\bigodot_{n\in[N]} \boldsymbol{\beta}_n^0 \right) \right\|_F^2 &= (1-a^N)^2 p \\
&\le (1-a^{2N}) p \\
&\le (1-a^2)(1+a^2+\cdots+a^{2(N-1)}) p \\
&\le \epsilon N p. \qquad (4.43)
\end{aligned}
$$

Moreover, for the second term in the bracket, $\sum_{\substack{i\in\{0,1\}^N \\ \|i\|_1\neq 0}} a^{2(N-\|\mathbf{i}\|_1)} b^{2\|\mathbf{i}\|_1} \left\| \bigodot_{n\in[N]} \boldsymbol{F}_n^{i_n} \right\|_F^2$,

we have

$$\begin{aligned}\sum_{\substack{i\in\{0,1\}^N\\ \|i\|_1\neq 0}} a^{2(N-\|\mathbf{i}\|_1)}b^{2\|\mathbf{i}\|_1}\left\|\bigodot_{n\in[N]}\boldsymbol{F}_n^{i_n}\right\|_F^2 &= \sum_{\substack{i\in\{0,1\}^N\\ \|i\|_1\neq 0}} a^{2(N-\|\mathbf{i}\|_1)}b^{2\|\mathbf{i}\|_1}p\\ &= p\sum_{\substack{i\in\{0,1\}^N\\ \|i\|_1\neq 0}}(1-\varepsilon)^{N-\|\mathbf{i}\|_1}\varepsilon^{\|\mathbf{i}\|_1}\\ &= p\sum_{N=0}^{n-1}\binom{n}{n}(1-\varepsilon)^n\varepsilon^{N-n}\\ &\overset{(a)}{=} p(1-(1-\varepsilon)^N)\\ &\overset{(b)}{\leq} \varepsilon Np.\end{aligned}$$

Thus, we have

$$\left\|\underline{\mathbf{B}}^t-\underline{\mathbf{B}}^0\right\|_F^2 \leq [2\varepsilon Np]\left\|\underline{\mathbf{G}}^*\right\|_F^2.$$

Therefore, by setting $\varepsilon \leq \frac{r^2}{2pN}$, we ensure that $\left\|\underline{\mathbf{B}}^t-\underline{\mathbf{B}}^0\right\|_F^2$, i.e., $\underline{\mathbf{B}}^t \in \mathcal{B}_r$. Remember that we also have $0<\varepsilon<1$ from (4.40). Therefore,

$$0<\varepsilon<\min\{1,\frac{r^2}{2pN}\}. \tag{4.44}$$

Lower bound on distance $\left\|\underline{\mathbf{B}}^t-\underline{\mathbf{B}}^0\right\|_F$: We now find a lower bound on the distance between any two elements in the set $\mathbb{B}_L$:

$$\begin{aligned}\left\|\underline{\mathbf{B}}^t-\underline{\mathbf{B}}^{t'}\right\|_F^2 &= \left\|\left(\bigotimes_{n\in[N]}\boldsymbol{\beta}_n^{t_n}-\bigotimes_{n\in[N]}\boldsymbol{\beta}_n^{t'_n}\right)\operatorname{vec}(\underline{\mathbf{G}}^*)\right\|_F^2\\ &= \operatorname{vec}(\underline{\mathbf{G}}^*)^T\left(\bigotimes_{n\in[N]}\boldsymbol{\beta}_n^{t_n}-\bigotimes_{n\in[N]}\boldsymbol{\beta}_n^{t'_n}\right)^T\left(\bigotimes_{n\in[N]}\boldsymbol{\beta}_n^{t_n}-\bigotimes_{n\in[N]}\boldsymbol{\beta}_n^{t'_n}\right)\operatorname{vec}(\underline{\mathbf{G}}^*)\\ &= \operatorname{vec}(\underline{\mathbf{G}}^*)^T\mathbf{M}\ \operatorname{vec}(\underline{\mathbf{G}}^*)\\ &\geq \sigma_{\min}^2(\mathbf{M})\left\|\operatorname{vec}(\underline{\mathbf{G}}^*)\right\|_2^2,\end{aligned} \tag{4.45}$$

where

$$\mathbf{M}_{t_n,t'_n} \triangleq \left(\bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t_n} - \bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t'_n}\right)^T \left(\bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t_n} - \bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t'_n}\right) \in \mathbb{R}^{p^N \times p^N},$$

and $\sigma_{\min}(\mathbf{M}_{t_n,t'_n})$ is the minimum singular value of $\mathbf{M}_{t_n,t'_n}$. Assuming that $\sigma_{\min} > 0$, we have

$$\sigma^2_{\min}(\mathbf{M}_{t_n,t'_n}) \geq \frac{\left\|\mathbf{M}_{t_n,t'_n}\right\|_F^2}{p^N\ \kappa^2_{t_n,t'_n}} \geq \frac{\left\|\mathbf{M}_{t_n,t'_n}\right\|_F^2}{p^N\ \kappa^2}, \tag{4.46}$$

where $\kappa_{t_n,t'_n} \triangleq \frac{\sigma^2_{\max}(\mathbf{M}_{t_n,t'_n})}{\sigma^2_{\min}(\mathbf{M}_{t_n,t'_n})}$ and $\kappa \triangleq \max_{t_n,t'_n} \kappa_{t_n,t'_n}$. In order to evaluate (4.46), we must find a lower bound on $\left\|\bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t_n} - \bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t'_n}\right\|_F^2$. We begin by stating that for any two distinct $\bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t_n}$ and $\bigotimes_{N\in[N]} \boldsymbol{\beta}_n^{t'_n}$, it is sufficient that $t_n \neq t'_n$ for only one $n \in [N]$ (only one factor matrix is different) [64]. Assume that N_d out of N factor matrices are distinct, and without loss of generality we assume that factor matrices labeled $1, \cdots, N_d$ are distinct, and factor matrices labeled $N_d + 1, \cdots, N$ are identical.

Thus we have:

$$
\begin{aligned}
\Big\| \bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t_n} - \bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t'_n} \Big\|_F^2 &= \Big\| \boldsymbol{\beta}_1^{t_1} \otimes \cdots \otimes \boldsymbol{\beta}_{N_d}^{t_{N_d}} \otimes \boldsymbol{\beta}_{N_d+1}^{t_{N_d+1}} \otimes \cdots \otimes \boldsymbol{\beta}_N^{t_N} \\
&\qquad - \boldsymbol{\beta}_1^{t'_1} \otimes \cdots \otimes \boldsymbol{\beta}_{N_d}^{t'_{N_d}} \otimes \boldsymbol{\beta}_{N_d+1}^{t'_{N_d+1}} \otimes \cdots \otimes \boldsymbol{\beta}_N^{t'_N} \Big\|_F^2 \\
&= \left\| \left(\bigotimes_{n=1}^{N_d} \boldsymbol{\beta}_n^{t_n} - \bigotimes_{n=1}^{N_d} \boldsymbol{\beta}_n^{t'_n} \right) \otimes \boldsymbol{\beta}_{N_d+1}^{t_{N_d+1}} \otimes \cdots \otimes \boldsymbol{\beta}_N^{t_N} \right\|_F^2 \\
&= \prod_{n=N_d+1}^{N} \|\boldsymbol{\beta}_n^{t_n}\|_F^2 \cdot \left\| \left(\bigotimes_{n=1}^{N_d} \boldsymbol{\beta}_n^{t_n} - \bigotimes_{n=1}^{N_d} \boldsymbol{\beta}_n^{t'_n} \right) \right\|_F^2 \\
&\overset{(a)}{=} p^{N-N_d} \left\| \sum_{\mathbf{i}\in\{0,1\}^{N_d}} a^{N_d - \|\mathbf{i}\|_1} b^{\|\mathbf{i}\|_1} \left(\bigotimes_{n=1}^{N_d} \boldsymbol{F}_n^{i_n} - \bigotimes_{n=1}^{N_d} \boldsymbol{F'}_n^{i_n} \right) \right\|_F^2 \\
&\overset{(b)}{=} p^{N-N_d} \cdot \sum_{\mathbf{i}\in\{0,1\}^{N_d},\ \|\mathbf{i}\|_1\neq 0} a^{2(N_d-\|\mathbf{i}\|_1)} b^{2\|\mathbf{i}\|_1} \\
&\qquad \cdot \prod_{\substack{n\in[N_d] \\ i_n=0}} \|\boldsymbol{\beta}_n^0\|_F^2 \cdot \left\| \bigotimes_{n\in[N_d]:\ i_n=1} \tilde{\boldsymbol{\beta}}_n^{t_n} - \bigotimes_{n\in[N_d]:\ i_n=1} \tilde{\boldsymbol{\beta}}_n^{t'_n} \right\|_F^2
\end{aligned}
\tag{4.47}
$$

where $\boldsymbol{F}_n^0 = \boldsymbol{F}_n'^0 = \boldsymbol{\beta}_n^0$, $\boldsymbol{F}_n^1 = \tilde{\boldsymbol{\beta}}_n^{t_n}$ and $\boldsymbol{F'}_n^1 = \tilde{\boldsymbol{\beta}}_n^{t'_n}$. Also, (a) follows from expansion (4.42), and (b) follows from orthogonality of the terms in the summation. For the term

$\left\|\bigotimes_{\substack{n\in[N]\\ i_n=1}}\tilde{\boldsymbol{\beta}}_n^{t_n}-\bigotimes_{\substack{k\in[K_d]\\ i_n=1}}\tilde{\boldsymbol{\beta}}_n^{t'_n}\right\|_F^2$ above, we have

$$\begin{aligned}
\left\|\bigotimes_{\substack{n\in[N]\\ i_n=1}}\tilde{\boldsymbol{\beta}}_n^{t_n}-\bigotimes_{\substack{n\in[N]\\ i_n=1}}\tilde{\boldsymbol{\beta}}_n^{t'_n}\right\|_F^2 &= \left\|\bigotimes_{\substack{n\in[N]\\ i_n=1}}\tilde{\boldsymbol{\beta}}_n^{t_n}\right\|_F^2+\left\|\bigotimes_{\substack{n\in[N]\\ i_n=1}}\tilde{\boldsymbol{\beta}}_n^{t'_n}\right\|_F^2-2\prod_{\substack{n\in[N]\\ i_n=1}}\langle\tilde{\boldsymbol{\beta}}_n^{t_n},\tilde{\boldsymbol{\beta}}_n^{t'_n}\rangle\\
&\overset{(a)}{\geq} 2\prod_{\substack{n\in[N_d]\\ i_n=1}}p-2\prod_{\substack{n\in[N_d]\\ i_n=1}}\left(\frac{1}{m_n-1}((m_n-1)p)-1)-\frac{1}{m_n-1}\right)\\
&= 2\prod_{\substack{n\in[N_d]\\ i_n=1}}p-2\prod_{\substack{n\in[N_d]\\ i_n=1}}\left(p-\frac{2}{m_n-1}\right)\\
&= 2\prod_{\substack{n\in[N_d]\\ i_n=1}}p-2\prod_{\substack{n\in[N_d]\\ i_n=1}}p\left(1-\frac{2}{(m_n-1)p}\right)\\
&\geq 2\prod_{\substack{n\in[N_d]\\ i_n=1}}p-2\prod_{\substack{n\in[N_d]\\ i_n=1}}p(1-t)\\
&= 2\left(\prod_{\substack{n\in[N_d]\\ i_n=1}}p-\prod_{\substack{n\in[N_d]\\ i_n=1}}p(1-t)\right)\\
&= 2\left(1-\prod_{\substack{n\in[N_d]\\ i_n=1}}(1-t)\right)\prod_{\substack{n\in[N_d]\\ i_n=1}}p\\
&\geq 2t\prod_{\substack{n\in[N_d]\\ i_n=1}}p, \qquad (4.48)
\end{aligned}$$

where $t\in(0,1)$ is such that $t\leq\min_{n\in[N]}\frac{2}{(m_n-1)p}$. and (a) follows from the fact that when $\bar{\boldsymbol{\beta}}_n^{t_n}$ and $\bar{\boldsymbol{\beta}}_n^{t'_n}$ differ in only one element, their inner product is greatest. Now, by

plugging (4.48) in (4.47), we get

$$\begin{aligned}
\Big\| \bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t_n} - \bigotimes_{n\in[N]} \boldsymbol{\beta}_n^{t'_n} \Big\|_F^2 &\geq 2tp^N \left[\sum_{\substack{i\in\{0,1\}^{N_d} \\ \|\mathbf{i}\|_1 \neq 0}} a^{2(N_d - \|\mathbf{i}\|_1)} b^{2\|\mathbf{i}\|_1} \right] \\
&= 2tp^N \sum_{n=0}^{N_d-1} \binom{N_d}{n} (1-\varepsilon)^n \varepsilon^{N_d-n} \\
&= 2tp^N \left[1 - (1-\varepsilon)^{N_d}\right] \\
&\geq 2tp^N \left[1 - (1-\varepsilon)\right] \\
&= 2tp^N \varepsilon. \qquad (4.49)
\end{aligned}$$

By replacing (4.49) in (4.45) and (4.46), we get

$$\begin{aligned}
\left\| \underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'} \right\|_F^2 &\geq \frac{2tp^N\varepsilon}{p^N\ \kappa^2} \|\underline{\mathbf{G}}\|_F^2 \\
&= \frac{2t\varepsilon}{\kappa^2} \|\underline{\mathbf{G}}\|_F^2. \qquad (4.50)
\end{aligned}$$

This means that the packing distance in this construction is $2\delta = \frac{2t\varepsilon}{\kappa^2_{t_n,t'_n}} \|\underline{\mathbf{G}}\|_F^2$.

Upper bounding mutual information: As stated in the problem formulation, the observations y follow a Normal distribution when conditioned on $\mathbf{X}$. Based on convexity of KL-divergence [142], we have that [139, 143]

$$\begin{aligned}
\mathbb{I}(t^*; y|\mathbf{X}) &= \frac{1}{T} \sum_{t\in T} \mathbb{E}_x \left[\mathbf{D}_{KL}\Big(f_{\underline{\mathbf{B}}^t}(\mathbf{y}|\mathbf{X}) || \frac{1}{T} \sum_{t'\in T} f_{\underline{\mathbf{B}}^{t'}}(\mathbf{y}|\mathbf{X}) \Big) \right] \\
&\leq \frac{1}{T^2} \sum_{t,t'\in T} \mathbb{E}_x \left[D_{KL}\Big(f_{\underline{\mathbf{B}}^t}(\mathbf{y}|\mathbf{X}) || f_{\underline{\mathbf{B}}^{t'}}(\mathbf{y}|\mathbf{X}) \Big) \right], \qquad (4.51)
\end{aligned}$$

where $D_{KL}(P_1||P_2)$ is the KL-divergence between two distributions P_1 and P_2, and $f_{\underline{\mathbf{B}}}(\mathbf{y}|X)$ is the probability distribution of responses $\mathbf{y}$ given coefficients tensor $\underline{\mathbf{B}}$ and

predictor tensors $\mathbf{X}$. For the KL-divergence. Since the conditional probability is Gaussian, from Durrieu et al. [144], we have

$$
\begin{aligned}
&\mathbb{E}_x D_{KL}(f_l(\mathbf{y}|\mathbf{X})||f_{l'}(\mathbf{y}|\mathbf{X})) \\
&\qquad = \mathbb{E}_x \left[\sum_{l=1}^{L} \frac{1}{2\sigma^2} \left\langle \underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'}, \underline{\mathbf{X}}_l \right\rangle^2 \right] \\
&\qquad = L\mathbb{E}_x \left[\frac{1}{2\sigma^2} \operatorname{vec}\left(\underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'} \right)^\top \operatorname{vec}(\underline{\mathbf{X}}_l) \operatorname{vec}(\underline{\mathbf{X}}_l)^\top \operatorname{vec}\left(\underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'} \right) \right] \\
&\qquad = L \left[\frac{1}{2\sigma^2} \operatorname{vec}\left(\underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'} \right)^\top \Sigma_x \operatorname{vec}\left(\underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'} \right) \right] \\
&\qquad \leq \frac{L}{2\sigma^2} \|\Sigma_x\|_2 \left\| \underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'} \right\|_2^2 . \qquad (4.52)
\end{aligned}
$$

It follows immediately from (4.52) that we must derive an upper bound on the distance between any two elements in the set $\mathbb{B}_T$:

$$
\begin{aligned}
\left\| \underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'} \right\|_F^2 &= \left\| \underline{\mathbf{B}}^t - \underline{\mathbf{B}}^0 + \underline{\mathbf{B}}^0 - \underline{\mathbf{B}}^{t'} \right\|_F^2 \\
&\leq \left\| \underline{\mathbf{B}}^t - \underline{\mathbf{B}}^0 \right\|_F^2 + \left\| \underline{\mathbf{B}}^0 - \underline{\mathbf{B}}^{t'} \right\|_F^2 \\
&\overset{(a)}{\leq} 4\varepsilon Np \|\underline{\mathbf{G}}^*\|_F^2, \qquad (4.53)
\end{aligned}
$$

where (a) follows from (4.44). Plugging in this upper bound for $\left\| \underline{\mathbf{B}}^t - \underline{\mathbf{B}}^{t'} \right\|_F^2$, we achieve the following upper bound for the mutual information in (4.51):

$$
\begin{aligned}
\mathbb{I}(t^*; \mathbf{y}|\mathbf{X}) &\leq \frac{1}{T^2} \sum_{t,t' \in T} \mathbb{E}_x \left[D_{KL}\left(f_{\underline{\mathbf{B}}^t}(\mathbf{y}|\mathbf{X}) || f_{\underline{\mathbf{B}}^{t'}}(\mathbf{y}|\mathbf{X}) \right) \right] \\
&\leq \frac{2LNp \|\underline{\mathbf{G}}^*\|_F^2 \|\Sigma_x\|_2}{\sigma^2} \varepsilon, \qquad (4.54)
\end{aligned}
$$

which concludes the proof of Lemma 15. □

Chapter 5

Momentum-based Accelerated Streaming PCA

5.1 Introduction

Principal component analysis (PCA) is a powerful tool with applications in machine learning, signal processing, and statistics. The aim in PCA is to learn directions of high variance (principal components) for a given dataset. This allows for representing the data using only the components (features) with highest variance and therefore reducing dimensionality of data while explaining as much variance in the data as possible. Reducing the dimensionality of data allows for more efficiently perform information processing and learning tasks especially when dealing with high dimensional data.

To find the principal components of a data matrix, one needs to find the top eigenvectors of the covariance matrix of data. Let the data samples be realizations of a random vector $\mathbf{x}$ generated from an unknown distribution $\mathcal{P}_x$ with zero mean and covariance matrix $\mathbf{\Sigma}$. The PCA problem can be posed as the statistical optimization problem

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times k}:\mathbf{W}^T\mathbf{W}=I} -\operatorname{Tr}(\mathbf{W}^T\mathbf{\Sigma}\mathbf{W}). \tag{5.1}$$

Since we often do not have access to $\mathcal{P}_x$ and therefore the true covariance matrix $\mathbf{\Sigma}$, we resort to solving the empirical PCA problem. That is, given N data samples $\mathbf{x}_n \in \mathbb{R}^d$, $n = 1, \cdots, N$ drawn from $\mathcal{P}_x$, we solve

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times k}:\mathbf{W}^T\mathbf{W}=I} -\frac{1}{N}\operatorname{Tr}(\mathbf{W}^T\mathbf{A}\mathbf{W}), \tag{5.2}$$

where $\mathbf{A} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T$ is the sample covariance matrix.

To solve this problem, iterative methods such as power method and Lanczos [10] are among the most popular methods. However, these methods need access to the sample covariance matrix $\mathbf{A}$ at every iteration. This is not possible in streaming settings where the algorithm observes data samples only one (or a few) at a time. Moreover, even in non-streaming settings, this requirement incurs an $O(d^2N)$ to compute the sample covariance matrix and an additional $O(d^2)$ memory for its storage, which can be prohibitive for machine learning applications where we typically work with large and high dimensional datasets.

To address these issues, streaming (stochastic) PCA algorithms such as Oja's rule [145, 146] and Krasulina's algorithm [147] have been proposed. These algorithms work with cheap-to-compute estimates of the covariance matrix in each iteration. That is, in the t-th iteration of these algorithms, an estimate $\mathbf{A}_t$ of the empirical covariance matrix is computed such that the algorithm does not need to access the entire data set in each iteration. This estimate is often chosen to be $\mathbf{A}_t = \mathbf{x}_t \mathbf{x}_t^T$ where $\mathbf{x}_t$ is the t-th observed data sample. In addition to being suitable for streaming settings and lower memory and computational cost compared to batch methods, these stochastic methods allow taking advantage of sparsity in data samples to reduce computational cost even further. In some applications such as natural language processing, computer vision, and recommendation systems we sometimes deal with data samples $\{\mathbf{x}_t\}$ and therefore estimates $\{\mathbf{A}_t\}$ that are sparse. However, this sparsity usually is not preserved in the sample covariance matrix $\mathbf{A} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{A}_t$.

Our focus in this work is on Oja's algorithm. Oja's simple update rule

$$\mathbf{w}_t' = \mathbf{w}_{t-1} + \eta_t \mathbf{A}_t \mathbf{w}_{t-1} \qquad \mathbf{w}_t = \mathbf{w}_t'/\|\mathbf{w}_t'\|, \tag{5.3}$$

where $\mathbf{A}_t = \mathbf{x}_t \mathbf{x}_t^T$, is perhaps the most popular streaming PCA algorithm. Oja's method can be seen as projected stochastic gradient descent (SGD) applied to the PCA problem (5.2). However, due to the nonconvexity of problem (5.2), the convergence guarantees for SGD do not directly apply here. Nonetheless, the convergence of Oja's method

to a global minimum of problem 5.2 is established in the literature. More precisely, the suboptimality error of Oja's and many other streaming PCA methods can be decomposed to a variance term (function of noise variance σ^2) and a bias term (function of initial error e_0). The bias component of suboptimality error is lower bounded as $\Omega(e_0 e^{-\sqrt{\lambda_1-\lambda_2}t})$ where λ_1 and λ_2 are respectively the largest and second largest eigenvalue of the sample covariance matrix [148]. On the other hand, the noise component of the error has a minimax lower bound $\Omega(\frac{\sigma^2}{(\lambda_1-\lambda_2)^2 t})$ [149]. To the best of our knowledge, the best convergence rate guarantees for streaming PCA are given by Jain et al. [150] where the authors show that with probability greater than 3/4, the iterates in Oja's method reach $O(\frac{\sigma^2}{(\lambda_1-\lambda_2)^2 t} + \frac{1}{t^2})$ error after t iterations.

Achieving such strong convergence results for the nonconvex PCA problem 5.2 is made possible perhaps due to the fact that this problem has a "nice" optimization landscape with escapable saddle points and no nonoptimal local minima [151–153]. This intuition encourages us to employ acceleration techniques that have been successfully implemented in many classes of convex optimization problems. In this chapter, inspired by recent works on accelerating stochastic gradient descent for certain classes of convex problems [154–157], we investigate whether a momentum-based acceleration method called the Polyak's heavy ball momentum method [14, 158] can help Oja's method achieve the lower bounds in both noiseless case (bias term) and noisy case (variance term).

We investigate different step size choices in the heavy ball accelerated Oja's method and propose a multi-stage scheme for choosing the step size. We prove the convergence of this multi-stage algorithm and show that, with high probability, it approximately (up to a log factor) achieves the $O\big(\frac{\sigma^2}{(\lambda_1-\lambda_2)^2 t}\big)$ upper bound in the bias term and $O(e_0 e^{-\sqrt{\lambda_1-\lambda_2}t})$ upper bound in the noise term. While the dependence of our convergence result on dimensions d is not optimal and there is an extra log factor in dependence on t, our results show that there could be benefit in applying momentum acceleration to stochastic solvers in this structured nonconvex problem. This result is also backed by our preliminary experimental results (see Figure 5.1).

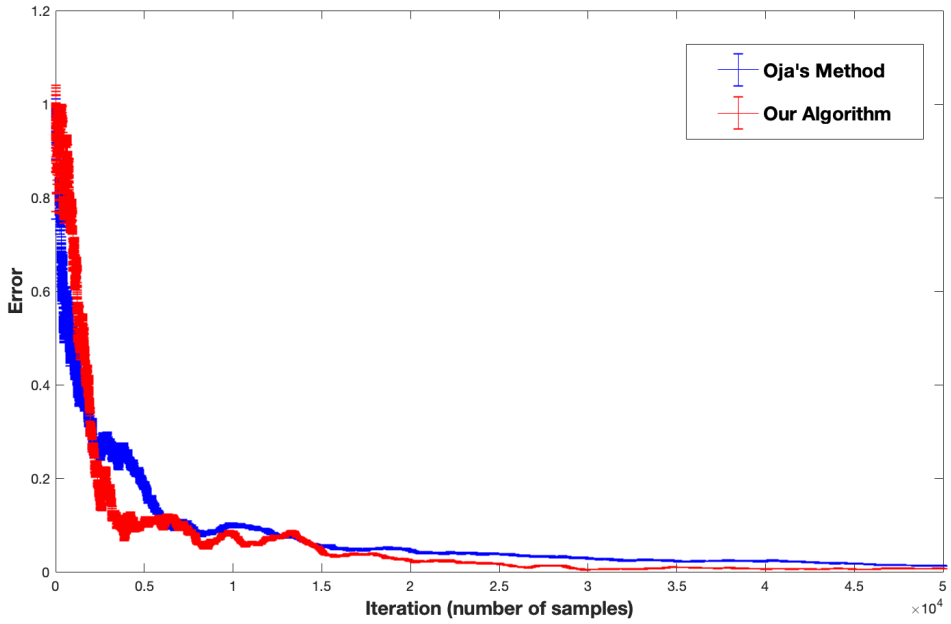


Figure 5.1: Performance of our proposed method and standard Oja's method in terms of error $(1 - \frac{\mathbf{u}_1^\top \mathbf{w}_t}{\|\mathbf{u}_1^\top \mathbf{w}_t\|_2})$ versus the number of iterations on a synthetic dataset. We can see improvement in the performance of our proposed heavy-ball momentum accelerated method compared to the standard non-accelerated method.

Remark. In this chapter, we focus on solving the 1-PCA problem. That is, computing the top eigenvector of a covariance matrix.

5.1.1 Relation to Prior Work

While earliest works on streaming PCA algorithms such as Oja's method [145, 146] and Krasulina's method [147] date back to 1980s, in the recent years there has been a renewed interest in streaming PCA methods due to widespread application of PCA in many machine learning and big data problems. Inspired by advances in obtaining non-asymptotic convergence rates for stochastic optimization, some works provide finite sample convergence results for classic streaming PCA algorithms, especially Oja's method [150, 159–162]. Many other works have focused on developing more efficient variants of these streaming PCA algorithms [148, 163–165]. Among these, Shamir [160], Xu et al. [148], and Kim and Klabjan [164] propose using variance reduction techniques to speed up the convergence. However, these algorithms are not suitable for true streaming settings since they require many passes over the data and also require

$O(d^2)$ memory compared to $O(d)$ memory cost of stochastic PCA methods without variance reduction. Moreover, Xu et al. [148], and Kim and Klabjan [164] , similar to our work, propose employing momentum-based acceleration to design stochastic PCA algorithms.

While the efficacy of acceleration methods such as Polyak's heavy ball method [158, 166] and Nesterov's accelerated gradient method [167] is well understood for deterministic (strongly) convex optimization problems, recently there has been a surge in interest in analyzing accelerated stochastic algorithms due to their scalability and their good performance in practice both in convex and nonconvex settings, especially in deep learning [155, 168–170]. Notably, Aybat et al. [155], Can et al. [170], and Jain et al. [154] study accelerated methods for strongly convex and certain classes of convex problems, while some other works [171–173] focus their attention to analyzing accelerated stochastic methods for nonconvex problems under mild conditions . While PCA is a nonconvex problem, the aforementioned results on nonconvex optimization only provide rates for first order convergence (convergence to a stationary point of the objective function) and not necessarily global convergence rates. In the 1-PCA problem which is the focus of this chapter, we are interested in finding the top eigenvector of the matrix which corresponds to global optima of Problem 5.2.

The idea of employing acceleration methods to speed up PCA algorithms has been proposed by Xu et al. [148], and Kim and Klabjan [164]. The proposed algorithms in these works however require working with large mini-batches or multiple passes over the data, making them undesirable for streaming settings. In contrast, we propose a heavy ball accelerated variant of Oja's method with multistage scheme for choosing the step size that is suitable for streaming settings since only requires access to a single data point per iteration. The multistage scheme for step size is adopted from Aybat et al. [155], where they show that heavy ball accelerated SGD with a similar multistage stepsize scheme achieves optimal convergence rate when applied to strongly convex, smooth functions.

5.2 Preliminaries and Problem Statement

5.2.1 Notation and Definitions

Throughout this write-up, scalars are represented by lower case letters: a, and vectors are denoted by boldface lower case letters: $\mathbf{a}$. Boldface upper case letters denote matrices: $\mathbf{A}$ and tensors are represented by boldface underlined upper case letters $\underline{\mathbf{A}}$.

We denote by $\|\mathbf{v}\|_p$ the ℓ_p norm of vector $\mathbf{v}$ (we abuse the terminology in case of $p = 0$), while we use $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_F$, and $\|\mathbf{A}\|_{\text{tr}}$ to denote the spectral, Frobenius, and trace (nuclear) norms of matrix $\mathbf{A}$, respectively.

We denote by $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{m_1 m_2 \times p_1 p_2}$ the Kronecker product of matrices $\mathbf{A} \in \mathbb{R}^{m_1 \times p_1}$ and $\mathbf{B} \in \mathbb{R}^{m_2 \times p_2}$. We use $\bigotimes_{n=1}^{N} \mathbf{A}_n \triangleq \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_N$ for the Kronecker product of N matrices. We drop the range indicators when there is no ambiguity.

5.2.2 The stochastic PCA problem

In streaming PCA, we want to find the top eigenvalue of a matrix $\mathbf{\Sigma}$, given a sequence of random samples $\mathbf{A}_t$ of $\mathbf{\Sigma}$ which are given to us in an online fashion. Here, we assume that the streaming algorithms have access to a stochastic oracle that provides (noisy unbiased) i.i.d. estimates $\mathbf{A}_t$ of a matrix $\mathbf{\Sigma}$ such that

$$\mathbb{E}\left[\mathbf{A}_t\right] = \mathbf{\Sigma}, \qquad \|\mathbf{A}_t\|_F \leq r, \qquad \mathbb{E}\left[\|\mathbf{A}_t - \mathbf{\Sigma}\|_F^2\right] = \sigma^2. \tag{5.4}$$

A special (and common) case of this setting is estimating the covariance matrix $\mathbf{\Sigma}$ by estimates $\mathbf{A}_t = \mathbf{x}_t \mathbf{x}_t^T$ where $\mathbf{x}_t$ is the random data sample presented to the algorithm at time t.

Further, we assume that the largest eigenvalue of $\mathbf{\Sigma}$ is strictly greater than the second largest eigenvalue, i.e., $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \lambda_d$.

5.2.3 Baseline Stochastic PCA Algorithms

The oldest and most well-known stochastic PCA algorithms are Oja's method [145, 146] and Krasulina's method [147]. In Oja's method, the update rule is

$$\mathbf{w}_t' = (\mathbf{I} + \eta \mathbf{A}_t)\mathbf{w}_{t-1}, \qquad \mathbf{w}_t = \mathbf{w}_t'/\|\mathbf{w}_t'\|, \tag{5.5}$$

which is sometimes written as

$$\mathbf{w}_t' = \frac{\mathbf{w}_{t-1} + \eta_t \mathbf{A}_t \mathbf{w}_{t-1}}{\|\mathbf{w}_{t-1} + \eta_t \mathbf{A}_t \mathbf{w}_{t-1}\|}, \qquad \mathbf{w}_t = \mathbf{w}_t'/\|\mathbf{w}_t'\|, \tag{5.6}$$

On the other hand, the update rule in Krasulina's algorithm is

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t \Big(\mathbf{A}_t - \frac{\mathbf{w}_{t-1}^\top \mathbf{A}_t \mathbf{w}_{t-1}}{\|\mathbf{w}_{t-1}\|^2} \mathbf{I} \Big) \mathbf{w}_{t-1}. \tag{5.7}$$

In this work, we focus on Oja's rule and its variants.

5.2.4 Momentum-based Acceleration of Gradient-based Optimization Methods

The baseline algorithm for solving a minimization problem with continuously differentiable objective function f is the gradient descent (GD) method with update rule

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \nabla f(\mathbf{w}_{t-1}), \tag{5.8}$$

where η_t is the stepsize (also known as learning rate in the machine learning community). The convergence rate of GD for convex functions with L-Lipschitz gradient is $O(L/\epsilon)$ and when the function is also μ-strongly convex, it is $O\big(L/\mu \log(1/\epsilon)\big)$ [157].

To improve the convergence of GD, accelerated first-order methods combine gradient information at the current and the past iterate, as well as the iterates themselves. Most common acceleration methods are Polyak's heavy ball momentum method [158, 166] and Nesterov's accelerated gradient method [167].

Polyak's method involves adding a "heavy-ball momentum" term $\beta(\mathbf{w}_{t-1} - \mathbf{w}_{t-2})$.

The update rule in the heavy ball accelerated GD is

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \beta(\mathbf{w}_{t-1} - \mathbf{w}_{t-2}) - \eta_t \nabla f(\mathbf{w}_{t-1}), \tag{5.9}$$

which sometimes is written in form of

$$\begin{aligned} \mathbf{v}_t &= \beta' \mathbf{v}_{t-1} - \nabla f(\mathbf{w}_{t-1}) \\ \mathbf{w}_t &= \mathbf{w}_{t-1} + \eta \mathbf{v}_t, \end{aligned} \tag{5.10}$$

where $\beta' = \frac{\beta'}{\eta_t}$. On the other hand, Nesterov [167] proposed a slightly different momentum method:

$$\begin{aligned} \mathbf{v}_t &= \beta' \mathbf{v}_{t-1} - \nabla_{\mathbf{w}} f(\mathbf{w}_{t-1} + \eta \beta' \mathbf{v}_{t-1}) \\ \mathbf{w}_t &= \mathbf{w}_{t-1} + \eta \mathbf{v}_t, \end{aligned} \tag{5.11}$$

which can also be written as

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \beta(\mathbf{w}_{t-1} - \mathbf{w}_{t-2}) - \eta \nabla_{\mathbf{w}} f(\mathbf{w}_{t-1} + \beta(\mathbf{w}_{t-1} - \mathbf{w}_{t-2})). \tag{5.12}$$

The heavy ball method has been shown to have $O\big(\sqrt{L/\mu}\log(1/\epsilon)\big)$ convergence rate when the objective function f is twice continuously differentiable, strongly convex and has Lipschitz continuous gradients, which is faster than both GD and Nesterov's accelerated gradient [157, 174]. However, the convergence of the heavy balling method is not as well-established when f is not necessarily twice differentiable [174, 175]. On the other hand, Nesterov's method has improved convergence rate for both classes of convex and strongly convex objective functions with Lipschitz continuous gradients. Especially, in the case of convex optimization problems with Lipschitz continuous gradients, Nesterov's method acheives the optimal rate of $O(\sqrt{L/\epsilon})$ [167, 174].

5.3 Oja's Rule with Heavy Ball Acceleration with Fixed Step Size

In this section we study the convergence of Oja's update rule with fixed step size and Polyak's heavy ball (HB) acceleration in the stochastic (streaming) setting. Consider the following variant of the Oja's rule:

$$\mathbf{w}_t' = (\mathbf{I} + \eta \mathbf{A}_t)\mathbf{w}_{t-1} - \beta \mathbf{w}_{t-2}, \qquad \mathbf{w}_t = \mathbf{w}_t'/\|\mathbf{w}_t'\|, \tag{5.13}$$

where A_t is the stochastic update at time t. We call this update rule the heavy ball accelerated Oja's Rule (HBOR), since the term $-\beta \mathbf{w}_{t-2}$ in the update works in similar way to the heavy ball momentum term.

Define random matrices $\mathbf{F}_t$ such that $\mathbf{w}_t' = \mathbf{F}_t \mathbf{w}_0$. Then (5.13) can be rewritten as a matrix recursion for $t \geq 1$:

$$\mathbf{F}_t = (\mathbf{I} + \eta \mathbf{A}_t)\mathbf{F}_{t-1} - \beta \mathbf{F}_{t-2}, \quad \mathbf{F}_0 = \mathbf{I}, \quad \mathbf{F}_{-1} = \mathbf{0}. \tag{5.14}$$

We can write update rule (5.13) in the compact form $\boldsymbol{\xi}_t' = \mathbf{M}_t \boldsymbol{\xi}_{t-1}$ where

$$\mathbf{M}_t \triangleq \begin{pmatrix} \mathbf{I} + \eta \mathbf{A}_t & -\beta \mathbf{I} \\ \mathbf{I} & 0 \end{pmatrix}, \tag{5.15}$$

and $\boldsymbol{\xi}_t = \begin{pmatrix} \mathbf{w}_t \\ \mathbf{w}_{t-1} \end{pmatrix}$ and $\boldsymbol{\xi}_t' = \begin{pmatrix} \mathbf{w}_t' \\ \mathbf{w}_{t-1}' \end{pmatrix}$. Since $\mathbf{w}_t' = \mathbf{F}_t \mathbf{w}_0$, one can write $\mathbf{F}_t = \mathbf{Z}^T \mathbf{M}_t \cdots \mathbf{M}_1 \mathbf{Z}$, where $\mathbf{Z} \triangleq \begin{pmatrix} I_d \\ \mathbf{0}_d \end{pmatrix}$.

Using the definition of $\mathbf{F}_t$ we have an expression for the residual error from projecting

the iterate on the top eigenvector:

$$\begin{aligned} 1 - \frac{\langle \mathbf{u}_1, \mathbf{w}_t' \rangle^2}{\|\mathbf{w}_t'\|^2} &= 1 - \frac{(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2}{\|\mathbf{F}_t \mathbf{w}_0\|^2} \\ &= \frac{\sum_{i=2}^d (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2}{\sum_{i=1}^d (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2} \\ &\leq \frac{\sum_{i=2}^d (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2}{(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2}. \end{aligned} \tag{5.16}$$

Thus, in order to find the convergence rate of the sequence generated by HBOR in (5.13), we first bound $\mathbb{E}\left[\sum_{i=2}^d (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2\right]$ and use Markov's inequality to obtain a high-probability upper bound on the numerator. For the denominator, we bound $\mathrm{Var}(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)$ which yields a lower bound on $|\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0|$ (and consequently $(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2$) using Chebyshev's inequality.

To establish an error bound of HB-Oja's rule with constant step size, we will first introduce a series of lemmata that are essential in obtaining our results.

Lemma 17. *Consider the update rule* (5.13) *and* $\mathbf{F}_t$ *defined in* (5.14)*. Assume that* $\|\mathbf{w}_0\| = 1$*. Then,*

$$\|\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right]\| \leq p_t^2(\lambda_1)\Big(\exp\Big[\frac{\sigma^2 \eta^2 t}{(1+\eta\lambda_1)^2 - 4\beta}\Big] - 1\Big), \tag{5.17}$$

where $\sigma^2 = \mathbb{E}\left[\|\mathbf{A} - \mathbf{A}_t\|^2\right]$*,* λ_1 *is the largest eigenvalue of matrix* $\mathbf{A}$*, and the polynomial sequence* $p_t(x)$ *is defined as*

$$p_t(x) = (1+\eta x)p_{t-1}(x) - \beta p_{t-2}(x), \quad p_1(x) = 1 + \eta x, \quad p_0(x) = 1. \tag{5.18}$$

Proof. We have $\mathbf{F}_t = \mathbf{Z}^T \mathbf{M}_t \cdots \mathbf{M}_1 \mathbf{Z}$. Therefore,

$$\begin{aligned} \mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] &= (\mathbf{Z} \otimes \mathbf{Z})^T \mathbb{E}\left[\mathbf{M}_t \otimes \mathbf{M}_t\right] \cdots \mathbb{E}\left[\mathbf{M}_1 \otimes \mathbf{M}_1\right] (\mathbf{Z} \otimes \mathbf{Z}) \\ &= (\mathbf{Z} \otimes \mathbf{Z})^T \mathbb{E}\left[\mathbf{M}_t \otimes \mathbf{M}_t\right]^t (\mathbf{Z} \otimes \mathbf{Z}). \end{aligned} \tag{5.19}$$

We have $\mathbf{M}_t = \mathbf{M} + \begin{pmatrix} \eta(\mathbf{A}_t - \mathbf{A}) & 0 \\ 0 & 0 \end{pmatrix}$. Define $\tilde{\mathbf{M}}_t \triangleq \mathbf{M} - \mathbf{M}_t$. Since $\mathbb{E}\left[\tilde{\mathbf{M}}_t\right] = 0$, we

have

$$\begin{aligned}\mathbb{E}\left[\mathbf{M}_t \otimes \mathbf{M}_t\right]^t &= \mathbb{E}\left[(\mathbf{M} - \tilde{\mathbf{M}}_t) \otimes (\mathbf{M} - \tilde{\mathbf{M}}_t)\right]^t \\ &= (\mathbf{M} \otimes \mathbf{M} + \mathbf{\Sigma})^t, \end{aligned} \tag{5.20}$$

where $\mathbf{\Sigma} \triangleq \mathbb{E}\left[\tilde{\mathbf{M}}_t \otimes \tilde{\mathbf{M}}_t\right]$. It is clear that

$$\mathbf{\Sigma} = \mathbb{E}\left[(\mathbf{Z}\eta(\mathbf{A}_t - \mathbf{A})\mathbf{Z}^T) \otimes (\mathbf{Z}\eta(\mathbf{A}_t - \mathbf{A})\mathbf{Z}^T)\right] = \eta^2(\mathbf{Z} \otimes \mathbf{Z})\mathbf{\Sigma}_{\mathbf{A}}(\mathbf{Z} \otimes \mathbf{Z})^T,$$

where $\mathbf{\Sigma}_{\mathbf{A}} \triangleq \mathbb{E}\left[(\mathbf{A}_t - \mathbf{A}) \otimes (\mathbf{A}_t - \mathbf{A})\right]$. Now, we know from the binomial expansion of matrices [148] that

$$(\mathbf{A} + \mathbf{B})^t = \sum_{n=0}^{t} \sum_{\mathbf{k} \in \mathbb{S}_{t-n}^{n+1}} \mathbf{A}^{k_1} \prod_{i=2}^{n+1} \mathbf{B}\mathbf{A}^{k_i},$$

where $\mathbb{S}_i^j \triangleq \{(k_1, \cdots, k_j) \in \mathbb{N}^j \mid k_1 + \cdots + k_j = i\}$. Applying this expansion to the right hand side of (5.20) results in

$$\begin{aligned}(\mathbf{M} \otimes \mathbf{M} + \mathbf{\Sigma})^t &= \sum_{n=0}^{t} \sum_{\mathbf{k} \in S_{t-n}^{n+1}} (\mathbf{M} \otimes \mathbf{M})^{k_1} \prod_{i=2}^{n+1} \mathbf{\Sigma}(\mathbf{M} \otimes \mathbf{M})^{k_i} \\ &= (\mathbf{M} \otimes \mathbf{M})^t + \sum_{n=1}^{t} \sum_{\mathbf{k} \in S_{t-n}^{n+1}} (\mathbf{M} \otimes \mathbf{M})^{k_1} \prod_{i=2}^{n+1} \mathbf{\Sigma}(\mathbf{M} \otimes \mathbf{M})^{k_i}.\end{aligned}$$

Since $\mathbb{E}\left[\mathbf{F}_t\right] = \mathbb{E}\left[\mathbf{Z}^T\mathbf{M}_t \cdots \mathbf{M}_1\mathbf{Z}\right] = \mathbf{Z}^T\mathbf{M}^t\mathbf{Z}$ and consequently $\mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right] = (\mathbf{Z} \otimes$

$\mathbf{Z})^T(\mathbf{M}\otimes\mathbf{M})^t(\mathbf{Z}\otimes\mathbf{Z})$, it follows that

$$
\begin{aligned}
&\|\mathbb{E}\left[\mathbf{F}_t\otimes\mathbf{F}_t\right]-\mathbb{E}\left[\mathbf{F}_t\right]\otimes\mathbb{E}\left[\mathbf{F}_t\right]\| \\
&=\|(\mathbf{Z}\otimes\mathbf{Z})^T\sum_{n=1}^{t}\sum_{\mathbf{k}}\left[(\mathbf{M}\otimes\mathbf{M})^{k_1}\prod_{i=2}^{n+1}\boldsymbol{\Sigma}(\mathbf{M}\otimes\mathbf{M})^{k_i}\right](\mathbf{Z}\otimes\mathbf{Z})\| \\
&=\|\sum_{n=1}^{t}\sum_{\mathbf{k}}(\mathbf{Z}\otimes\mathbf{Z})^T\left[(\mathbf{M}\otimes\mathbf{M})^{k_1}\prod_{i=2}^{n+1}\eta^2(\mathbf{Z}\otimes\mathbf{Z})\cdot\boldsymbol{\Sigma}_{\mathbf{A}}\cdot(\mathbf{Z}\otimes\mathbf{Z})^T(\mathbf{M}\otimes\mathbf{M})^{k_i}\right](\mathbf{Z}\otimes\mathbf{Z})\| \\
&=\|\sum_{n=1}^{t}\sum_{\mathbf{k}}(\mathbf{Z}\otimes\mathbf{Z})^T(\mathbf{M}\otimes\mathbf{M})^{k_1}(\mathbf{Z}\otimes\mathbf{Z})\prod_{i=2}^{n+1}\left[\eta^2\boldsymbol{\Sigma}_{\mathbf{A}}(\mathbf{Z}\otimes\mathbf{Z})^T(\mathbf{M}\otimes\mathbf{M})^{k_i}(\mathbf{Z}\otimes\mathbf{Z})\right]\| \\
&=\|\sum_{n=1}^{t}\sum_{\mathbf{k}}(\mathbb{E}\left[\mathbf{F}_{k_1}\right]\otimes\mathbb{E}\left[\mathbf{F}_{k_1}\right])\prod_{i=2}^{n+1}\left[\eta^2\boldsymbol{\Sigma}_{\mathbf{A}}(\mathbb{E}\left[\mathbf{F}_{k_i}\right]\otimes\mathbb{E}\left[\mathbf{F}_{k_i}\right])\right]\| \\
&\leq\sum_{n=1}^{t}\sum_{\mathbf{k}}\|(\mathbb{E}\left[\mathbf{F}_{k_1}\right]\otimes\mathbb{E}\left[\mathbf{F}_{k_1}\right])\prod_{i=2}^{n+1}\left[\eta^2\boldsymbol{\Sigma}_{\mathbf{A}}(\mathbb{E}\left[\mathbf{F}_{k_i}\right]\otimes\mathbb{E}\left[\mathbf{F}_{k_i}\right])\right]\| \\
&\leq\sum_{n=1}^{t}\sum_{\mathbf{k}\in\mathbb{S}_{t-n}^{n+1}}\|\mathbb{E}\left[\mathbf{F}_{k_1}\right]\otimes\mathbb{E}\left[\mathbf{F}_{k_1}\right]\|_2\cdot\|\prod_{i=2}^{n+1}\eta^2\|\boldsymbol{\Sigma}_{\mathbf{A}}\|_2\cdot\|\mathbb{E}\left[\mathbf{F}_{k_i}\right]\otimes\mathbb{E}\left[\mathbf{F}_{k_i}\right]\|_2 \\
&\leq\sum_{n=1}^{t}\|\eta^2\boldsymbol{\Sigma}_{\mathbf{A}}\|^n\sum_{\mathbf{k}\in\mathbb{S}_{t-n}^{n+1}}\|\mathbb{E}\left[\mathbf{F}_{k_1}\right]\otimes\mathbb{E}\left[\mathbf{F}_{k_1}\right]\|_2\prod_{i=2}^{n+1}\|\mathbb{E}\left[\mathbf{F}_{k_i}\right]\otimes\mathbb{E}\left[\mathbf{F}_{k_i}\right]\|_2 \\
&=\sum_{n=1}^{t}\|\eta^2\boldsymbol{\Sigma}_{\mathbf{A}}\|^n\sum_{\mathbf{k}\in\mathbb{S}_{t-n}^{n+1}}\prod_{i=1}^{n+1}\|\mathbb{E}\left[\mathbf{F}_{k_i}\right]\otimes\mathbb{E}\left[\mathbf{F}_{k_i}\right]\|_2. \qquad (5.21)
\end{aligned}
$$

Now, using the properties of the polynomial sequence $p_t(x)$ described by (5.18) and the

fact that $\|\mathbb{E}\left[\mathbf{F}_{k_i}\right]\|_2 = \|p_{k_i}(\mathbf{A})\|_2 = p_{k_i}(\lambda_1)$, we get

$$
\begin{aligned}
&\|\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right]\| \\
&\qquad \leq \sum_{n=1}^{t} \|\eta^2 \mathbf{\Sigma_A}\|^n \sum_{\mathbf{k} \in \mathbb{S}_{t-n}^{n+1}} \prod_{i=1}^{n+1} \|\mathbb{E}\left[\mathbf{F}_{k_i}\right]\|_2^2 \\
&\qquad = \sum_{n=1}^{t} \eta^{2n} \|\mathbf{\Sigma_A}\|^n \sum_{\mathbf{k} \in \mathbb{S}_{t-n}^{n+1}} \prod_{i=1}^{n+1} p_{k_i}^2(\lambda_1) \\
&\qquad \overset{(a)}{\leq} \sum_{n=1}^{t} \eta^{2n} \|\mathbf{\Sigma_A}\|^n \sum_{\mathbf{k} \in \mathbb{S}_{t-n}^{n+1}} \frac{1}{((1+\eta\lambda_1)^2 - 4\beta)^n} p_{n+\sum_{i=1}^{n+1} k_i}^2(\lambda_1) \\
&\qquad \overset{(b)}{=} p_t^2(\lambda_1) \sum_{n=1}^{t} \eta^{2n} \|\mathbf{\Sigma_A}\|^n \binom{t}{t-n} \frac{1}{((1+\eta\lambda_1)^2 - 4\beta)^n} \\
&\qquad = p_t^2(\lambda_1) \sum_{n=1}^{t} \binom{t}{t-n} \Big(\frac{\eta^2 \|\mathbf{\Sigma_A}\|}{(1+\eta\lambda_1)^2 - 4\beta}\Big)^n \\
&\qquad \overset{(c)}{=} p_t^2(\lambda_1) \Big(\Big[\frac{\eta^2 \|\mathbf{\Sigma_A}\|}{(1+\eta\lambda_1)^2 - 4\beta} + 1\Big]^t - 1 \Big) \\
&\qquad \overset{(d)}{\leq} p_t^2(\lambda_1) \Big(\exp\Big[\frac{\eta^2 \|\mathbf{\Sigma_A}\| t}{(1+\eta\lambda_1)^2 - 4\beta}\Big] - 1 \Big), \qquad (5.22)
\end{aligned}
$$

where (a) follows from Corollary 4 in the Appendix (Section 6.2), (b) follows from $|\mathbb{S}_{t-n}^{n+1}| = \binom{t}{t-n}$, (c) follows from the binomial theorem [148], and (d) follows from $1 + x \leq \exp(x)$. Finally, since $||\mathbf{\Sigma_A}|| = \mathbb{E}\left[\|\mathbf{A} - \mathbf{A}_t\|^2\right] = \sigma^2$, we have

$$
\|\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right]\| \leq p_t^2(\lambda_1) \Big(\exp\Big[\frac{\sigma^2 \eta^2 t}{(1+\eta\lambda_1)^2 - 4\beta}\Big] - 1 \Big). \qquad (5.23)
$$

□

The following lemma provides a high probability upper bound on the numerator of the right hand side of inequality (5.16), i.e., $\sum_{i=2}^{d} (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2$.

Lemma 18. *Consider the update rule* (5.13) *and* $\mathbf{F}_t$ *defined in* (5.14). *Assume that*

$\|\mathbf{w}_0\| = 1$. *Then, with probability at least* $1-\delta$,

$$\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2 \leq \frac{1}{\delta}\left[\sqrt{d}\; p_t^2(\lambda_1)\big(\exp\Big[\frac{\sigma^2\eta^2 t}{(1+\eta\lambda_1)^2-4\beta}\Big]-1\big) + p_t^2(\lambda_2)(1-|\mathbf{u}_1^T\mathbf{w}_0|^2)\right]. \tag{5.24}$$

Proof of Lemma 18. To find an upper bound on the numerator $\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2$, we use Markov's inequality [45, 141]. To avoid complications of finding the exact value of $\mathbb{E}\left[\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2\right]$, we resort to finding an upper bound on it. We have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2\right]\\
&=\sum_{i=2}^{d}(\mathbb{E}\left[(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2\right]-\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right]^2)+\sum_{i=2}^{d}\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right]^2\\
&\leq\sum_{i=1}^{d}(\mathbb{E}\left[(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2\right]-\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right]^2)+\sum_{i=2}^{d}\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right]^2\\
&=\sum_{i=1}^{d}(\mathbb{E}\left[(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)\otimes(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)\right]-\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right]\otimes\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right])+\sum_{i=2}^{d}\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right]^2\\
&=\sum_{i=1}^{d}(\mathbf{u}_i\otimes\mathbf{u}_i)^T(\mathbb{E}\left[\mathbf{F}_t\otimes\mathbf{F}_t\right]-\mathbb{E}\left[\mathbf{F}_t\right]\otimes\mathbb{E}\left[\mathbf{F}_t\right])(\mathbf{w}_0\otimes\mathbf{w}_0)+\sum_{i=2}^{d}\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right]^2\\
&\overset{(a)}{\leq}\left\|\sum_{i=1}^{d}(\mathbf{u}_i\otimes\mathbf{u}_i)\right\|\cdot\|\mathbb{E}\left[\mathbf{F}_t\otimes\mathbf{F}_t\right]-\mathbb{E}\left[\mathbf{F}_t\right]\otimes\mathbb{E}\left[\mathbf{F}_t\right]\|\cdot\|\mathbf{w}_0\otimes\mathbf{w}_0\|+\sum_{i=2}^{d}(p_t(\lambda_i)\mathbf{u}_i^T\mathbf{w}_0)^2\\
&\leq\sqrt{d}\; p_t^2(\lambda_1)\big(\exp\Big[\frac{\sigma^2\eta^2 t}{(1+\eta\lambda_1)^2-4\beta}\Big]-1\big)+p_t^2(\lambda_2)\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{w}_0)^2\\
&\overset{(b)}{=}\sqrt{d}\; p_t^2(\lambda_1)\big(\exp\Big[\frac{\sigma^2\eta^2 t}{(1+\eta\lambda_1)^2-4\beta}\Big]-1\big)+p_t^2(\lambda_2)(1-|\mathbf{u}_1^T\mathbf{w}_0|^2)
\end{aligned}\tag{5.25}$$

where (a) follows from Cauchy–Schwarz inequality, (b) follows from orthonormality of $\mathbf{u}_i$'s as basis vector. Using Markov's inequality, we have that with probability at least $1-\delta$,

$$\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2 \leq \frac{1}{\delta}\left[\sqrt{d}\; p_t^2(\lambda_1)\big(\exp\Big[\frac{\sigma^2\eta^2 t}{(1+\eta\lambda_1)^2-4\beta}\Big]-1\big) + p_t^2(\lambda_2)(1-|\mathbf{u}_1^T\mathbf{w}_0|^2)\right]$$

for any fixed $0<\delta<1$. □

Next, the following lemma provides a high probability lower bound on the denominator of the right hand side of inequality (5.16), i.e., $(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2$.

Lemma 19. *Consider the update rule* (5.13) *and* $\mathbf{F}_t$ *defined in* (5.14). *Assume that* $\|\mathbf{w}_0\| = 1$. *Then, with probability at least* $1 - \delta$ *we have*

$$(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2 \geq p_t^2(\lambda_1) \left[|\mathbf{u}_1^T \mathbf{w}_0| - \sqrt{\frac{\exp\left[\frac{\sigma^2 \eta^2 t}{(1+\eta\lambda_1)^2 - 4\beta}\right] - 1}{\delta}} \right]^2. \tag{5.26}$$

Proof of Lemma 19. We use Chebyshev's inequality [45, 141, 176] to find a lower bound on the value of $(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2$. The variance of the denominator $\mathrm{Var}(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)$ thus is bounded as follows.

$$\begin{aligned}
\mathrm{Var}(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0) &= \mathbb{E}\left[(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2\right] - \left(\mathbb{E}\left[(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)\right]\right)^2 \\
&= \mathbb{E}\left[(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0) \otimes (\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)\right] - \mathbb{E}\left[(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)\right] \otimes \mathbb{E}\left[(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)\right] \\
&= (\mathbf{u}_1 \otimes \mathbf{u}_1)^T (\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right])(\mathbf{w}_0 \otimes \mathbf{w}_0) \\
&\leq \|\mathbf{u}_1 \otimes \mathbf{u}_1\| \cdot \|\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right]\| \cdot \|\mathbf{w}_0 \otimes \mathbf{w}_0\| \\
&\leq p_t^2(\lambda_1) \Big(\exp\Big[\frac{\sigma^2 \eta^2 t}{(1+\eta\lambda_1)^2 - 4\beta}\Big] - 1 \Big).
\end{aligned} \tag{5.27}$$

Using Chebyshev's inequality we get

$$\mathbb{P}\left(|\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0 - \mathbf{u}_1^T p_t(\mathbf{A}) \mathbf{w}_0| \geq p_t(\lambda_1) \frac{\sqrt{\exp\left[\frac{\sigma^2 \eta^2 t}{(1+\eta\lambda_1)^2 - 4\beta}\right] - 1}}{\sqrt{\delta}} \right) \leq \delta.$$

Note that

$$\mathbf{u}_i^T p_t(\mathbf{A}) = \mathbf{u}_i^T p_t(\lambda_i) \mathbf{u}_i \mathbf{u}_i^T = p_t(\lambda_i) \mathbf{u}_i^T. \tag{5.28}$$

Therefore, we have

$$\mathbb{P}\left(|\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0 - p_t(\lambda_1) \mathbf{u}_1^T \mathbf{w}_0| \geq p_t(\lambda_1) \frac{\sqrt{\exp\left[\frac{\sigma^2 \eta^2 t}{(1+\eta\lambda_1)^2 - 4\beta}\right] - 1}}{\sqrt{\delta}} \right) \leq \delta. \tag{5.29}$$

Thus,

$$\mathbb{P}\left(|\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0| \leq p_t(\lambda_1)|\mathbf{u}_1^T\mathbf{w}_0| - p_t(\lambda_1)\frac{\sqrt{\exp\left[\frac{\sigma^2\eta^2 t}{(1+\eta\lambda_1)^2-4\beta}\right]-1}}{\sqrt{\delta}}\right) \leq \delta \tag{5.30}$$

and consequently

$$\mathbb{P}\left((\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0)^2 \leq p_t^2(\lambda_1)\left[|\mathbf{u}_1^T\mathbf{w}_0| - \sqrt{\frac{\exp\left[\frac{\sigma^2\eta^2 t}{(1+\eta\lambda_1)^2-4\beta}\right]-1}{\delta}}\right]^2\right) \leq \delta. \tag{5.31}$$

□

Next, we investigate the term $\frac{p_t^2(\lambda_2)}{p_t^2(\lambda_1)}$ that will appear in the upper bound on the error. The following lemma states the result.

Lemma 20. *Given the polynomial sequence $\{p_t(x)\}$ defined as*

$$p_t(x) = (1+\eta x)p_{t-1}(x) - \beta p_{t-2}(x), \quad p_1(x) = 1+\eta x, \quad p_0(x) = 1,$$

and $\beta = (1+\eta\lambda_2)^2/4$, we have

$$\frac{p_t^2(\lambda_2)}{p_t^2(\lambda_1)} \leq \Big(\frac{t+1}{\sum_{n=0}^{t} a^{-2n}}\Big)^2 \exp\Big(-2t\frac{\sqrt{\eta\Delta}}{\sqrt{1+\eta\lambda_2}}\Big), \tag{5.32}$$

where $a \triangleq \frac{\mu_{1+}}{\mu_2} = \frac{(1+\eta\lambda_1)+\sqrt{(1+\eta\lambda_1)^2-(1+\eta\lambda_2)^2}}{1+\eta\lambda_2}$ and $\Delta = \lambda_1 - \lambda_2$.

Proof of Lemma 20. It follows from Lemma 22 that when $y \triangleq 1+\eta x$ is such that $y^2 \neq 4\beta$, we have

$$p_t(x) = \frac{1}{\sqrt{y^2-4\beta}}\left[\left(\frac{y+\sqrt{y^2-4\beta}}{2}\right)^{t+1} - \left(\frac{y-\sqrt{y^2-4\beta}}{2}\right)^{t+1}\right],$$

and when $y^2 = 4\beta$, we have

$$p_t(x) = (t+1)(\sqrt{\beta})^t. \tag{5.33}$$

Plugging in our choice of $\beta = (1+\eta\lambda_2)^2/4$ results in

$$\begin{aligned} p_t(\lambda_1) &= \frac{\left(\frac{(1+\eta\lambda_1)+\sqrt{(1+\eta\lambda_1)^2-(1+\eta\lambda_2)^2}}{2}\right)^{t+1} - \left(\frac{(1+\eta\lambda_1)-\sqrt{(1+\eta\lambda_1)^2-(1+\eta\lambda_2)^2}}{2}\right)^{t+1}}{\sqrt{(1+\eta\lambda_1)^2-(1+\eta\lambda_2)^2}} \\ &= \frac{\mu_{1_+}^{t+1} - \mu_{1_-}^{t+1}}{\mu_{1_+} - \mu_{1_-}} \end{aligned} \tag{5.34}$$

and

$$p_t(\lambda_2) = (t+1)((1+\eta\lambda_2)/2)^t = (t+1)\big(\frac{\mu_2}{2}\big)^t. \tag{5.35}$$

$$\begin{aligned} \frac{p_t(\lambda_2)}{p_t(\lambda_1)} &= \frac{(\mu_{1_+} - \mu_{1_-})(t+1)\mu_2^t}{\mu_{1_+}^{t+1} - \mu_{1_-}^{t+1}} \\ &= \frac{(\mu_{1_+} - \mu_{1_-})(t+1)\mu_2^t}{(\mu_{1_+} - \mu_{1_-})\sum_{n=0}^{t}\mu_{1_+}^{t-n}\mu_{1_-}^{n}}. \end{aligned} \tag{5.36}$$

Since $\mu_{1_+}\mu_{1_-} = (1+\eta\lambda_2)^2/4 = \mu_2^2$, we have $\frac{\mu_{1_+}}{\mu_2} = \frac{\mu_2}{\mu_{1_-}}$. Define $a \triangleq \frac{\mu_{1_+}}{\mu_2}$. We have

$$\begin{aligned} \frac{p_t(\lambda_2)}{p_t(\lambda_1)} &= (t+1)\frac{\mu_2^t}{\sum_{n=0}^{t}\mu_{1_+}^{t-n}\mu_{1_-}^{n}} \\ &= (t+1)\frac{\mu_2^t}{\sum_{n=0}^{t}a^{t-n}\mu_2^{t-n}a^{-n}\mu_2^{n}} \\ &= (t+1)\frac{1}{\sum_{n=0}^{t}a^{t-n}(\frac{1}{a})^{n}} \\ &= (t+1)\frac{1}{a^t\sum_{n=0}^{t}a^{-2n}}. \end{aligned} \tag{5.37}$$

Now, by showing that $\sum_{n=0}^{t}a^{t-2n}$ attains its minimum over $a \geq 1$ at $a = 1$, we show that $\frac{p_t(\lambda_2)}{p_t(\lambda_1)} < 1$ for all t. We have

$$\begin{aligned} \frac{\partial}{\partial a}\sum_{n=0}^{t}a^{t-2n} &= \sum_{n=0}^{t}(t-2n)a^{t-2n-1} \\ &= \sum_{n=0}^{\lfloor t/2\rfloor}(t-2n)a^{t-2n-1} + \sum_{n=\lfloor t/2\rfloor+1}^{t}(t-2n)a^{t-2n-1}. \end{aligned} \tag{5.38}$$

It is clear that the first term has non-negative multiplier and exponent while the second term has non-positive multiplier and exponent which make both terms increasing in a for $a \geq 1$. Since $\frac{\partial}{\partial a}\sum_{n=0}^{t} a^{t-2n} = 0$ for $a = 1$, we have $\frac{\partial}{\partial a}\sum_{n=0}^{t} a^{t-2n} \geq 1$ for $a > 1$ which means $\sum_{n=0}^{t} a^{t-2n} > t+1$ for all t and $a > 1$. Therefore,

$$\frac{p_t(\lambda_2)}{p_t(\lambda_1)} = (t+1)\frac{1}{a^t \sum_{n=0}^{t} a^{-2n}} = \frac{t+1}{\sum_{n=0}^{t} a^{-2n}}(\frac{\mu_2}{\mu_{1_+}})^t < 1. \tag{5.39}$$

We further have

$$\begin{aligned}
(\frac{\mu_2}{\mu_{1_+}})^{2t} &= \exp\left(2t\log(|\frac{\mu_{2_+}}{\mu_{1_+}}|)\right) \\
&= \exp\left(2t\log(\frac{\sqrt{4\beta}}{(1+\eta\lambda_1)+\sqrt{(1+\eta\lambda_1)^2-4\beta}})\right) \\
&\overset{(1)}{=} \exp\left(2t\log(\frac{1+\eta\lambda_2}{(1+\eta\lambda_2)+\eta(\lambda_1-\lambda_2)+\sqrt{(1+\eta\lambda_1)^2-(1+\eta\lambda_2)^2}})\right) \\
&\leq \exp\left(-2t\log(1+\frac{\eta(\lambda_1-\lambda_2)+\sqrt{(1+\eta\lambda_1)^2-(1+\eta\lambda_2)^2}}{1+\eta\lambda_2})\right) \\
&\overset{(2)}{\leq} \exp\left(-2t\frac{\eta(\lambda_1-\lambda_2)+\sqrt{(1+\eta\lambda_1)^2-(1+\eta\lambda_2)^2}}{1+\eta\lambda_2}\right) \\
&= \exp\left(-2t\frac{\eta(\lambda_1-\lambda_2)+\sqrt{(1+\eta\lambda_1)+(1+\eta\lambda_2)}\sqrt{(1+\eta\lambda_1)-(1+\eta\lambda_2)}}{1+\eta\lambda_2}\right) \\
&= \exp\left(-2t\frac{\eta(\lambda_1-\lambda_2)+\sqrt{(1+\eta\lambda_1)+(1+\eta\lambda_2)}\sqrt{\eta(\lambda_1-\lambda_2)}}{1+\eta\lambda_2}\right) \\
&= \exp\left(-2t\frac{\sqrt{1+\eta\lambda_2}\sqrt{\eta(\lambda_1-\lambda_2)}}{1+\eta\lambda_2}\right)\exp\left(-2t\frac{\eta(\lambda_1-\lambda_2)}{1+\eta\lambda_2}\right) \\
&\leq \exp\left(-2t\frac{\sqrt{\eta\Delta}}{\sqrt{1+\eta\lambda_2}}\right),
\end{aligned} \tag{5.40}$$

where $\Delta = \lambda_1 - \lambda$ is the eigengap of matrix $\mathbf{A}$. Hence, we have

$$\frac{p_t^2(\lambda_2)}{p_t^2(\lambda_1)} \leq \Big(\frac{t+1}{\sum_{n=0}^{t} a^{-2n}}\Big)^2 \exp\Big(-2t\frac{\sqrt{\eta\Delta}}{\sqrt{1+\eta\lambda_2}}\Big). \tag{5.41}$$

□

Now we are ready to provide an error bound on HBOR with constant step size.

Theorem 9 states this result.

Theorem 9. *Consider a PSD matrix* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *with eigenvalues* $1 \geq \lambda_1 \geq \cdots \geq \lambda_d$ *and eigengap* $\Delta := \lambda_1 - \lambda_2$. *Assume that estimates* $\mathbf{A}_t$ *of matrix* $\mathbf{A}$ *satisfy the set of assumptions in* (5.4). *Consider the estimates* $\mathbf{w}_t$ *generated by the accelerated Oja's method with heavy-balling*

$$\mathbf{w}'_t = (\mathbf{I} + \eta\mathbf{A}_t)\mathbf{w}_{t-1} - \beta\mathbf{w}_{t-2}, \qquad \mathbf{w}_t = \mathbf{w}'_t/\|\mathbf{w}'_t\|,$$

with $\beta = (1+\eta\lambda_2)^2/4$ *and constant stepsize* η. *Suppose that for some* $\iota > 0$ *we have* $\eta\sigma^2 < \frac{(2+\eta(\lambda_1+\lambda_2))\cdot\Delta\cdot(|\mathbf{u}_1^T\mathbf{w}_0|-\sqrt{\iota})^2}{T}$. *Then, with probability at least* $1-2\delta$, *we have*

$$e_T \triangleq 1 - \frac{\langle \mathbf{u}_1, \mathbf{w}'_T\rangle^2}{\|\mathbf{w}'_T\|^2} \leq K_1\eta T + K_2 e_0 \Big(\frac{T+1}{\sum_{t=0}^T a^{-2t}}\Big)^2 e^{-\gamma\sqrt{\eta}T}, \tag{5.42}$$

where $K_1 = \frac{\sqrt{d}2\sigma^2}{\delta(2+\eta(\lambda_1+\lambda_2))\Delta\iota}$, $K_2 = \frac{1}{\delta\iota}$, $a = \frac{(1+\eta\lambda_1)+\sqrt{(1+\eta\lambda_1)^2-(1+\eta\lambda_2)^2}}{1+\eta\lambda_2}$, *and* $\gamma = 2\frac{\sqrt{\Delta}}{\sqrt{1+\eta\lambda_2}}$.

Proof of Theorem 9. We want a high probability upper bound on (5.16) to get an upper bound on $1 - \frac{\langle \mathbf{u}_1, \mathbf{w}'_t\rangle^2}{\|\mathbf{w}'_t\|^2}$. We saw in inequality (5.16) that $1 - \frac{\langle \mathbf{u}_1, \mathbf{w}'_t\rangle^2}{\|\mathbf{w}'_t\|^2} \leq \frac{\sum_{i=2}^d (\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2}{(\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0)^2}$. Now, we find a high probability upper bound on the numerator of the right hand side, $\sum_{i=2}^d (\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2$, and a high probability lower bound on the denominator, $(\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0)^2$. We showed in Lemma 18 that with probability at least $1-\delta$,

$$\sum_{i=2}^d (\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2 \leq \frac{1}{\delta}\left[\sqrt{d}\; p_t^2(\lambda_1)\big(\exp\Big[\frac{\sigma^2\eta^2 t}{(1+\eta\lambda_1)^2-4\beta}\Big] - 1\big) + p_t^2(\lambda_2)(1-|\mathbf{u}_1^T\mathbf{w}_0|^2)\right].$$

Furthermore, we showed in Lemma 19 that with probability not less than $1-\delta$,

$$(\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0)^2 \geq p_t^2(\lambda_1)\left[|\mathbf{u}_1^T\mathbf{w}_0| - \sqrt{\frac{\exp\big[\frac{\sigma^2\eta^2 t}{(1+\eta\lambda_1)^2-4\beta}\big]-1}{\delta}}\right]^2.$$

It follows from these results, using a union bound, that with probability at least $1-2\delta$,

$$\frac{\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2}{(\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0)^2} \geq \frac{\sqrt{d}\; p_t^2(\lambda_1)\big(\exp\left[\frac{\sigma^2\eta^2 t}{(1+\eta\lambda_1)^2-4\beta}\right]-1\big)+p_t^2(\lambda_2)(1-|\mathbf{u}_1^T\mathbf{w}_0|^2)}{\delta p_t^2(\lambda_1)\left[|\mathbf{u}_1^T\mathbf{w}_0|-\sqrt{(\exp\left[\frac{\sigma^2\eta^2 t}{(1+\eta\lambda_1)^2-4\beta}\right]-1)/\delta}\right]^2}. \quad (5.43)$$

Plugging in $\beta=(1+\eta\lambda_2)^2/4$, we get with probability at least $1-2\delta$,

$$\begin{aligned}\frac{\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2}{(\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0)^2} \geq & \frac{\sqrt{d}\left(\exp\left[\frac{\sigma^2\eta t}{(2+\eta(\lambda_1+\lambda_2))\Delta}\right]-1\right)}{\delta\Big[|\mathbf{u}_1^T\mathbf{w}_0|-\sqrt{\frac{\exp\left[\frac{\sigma^2\eta t}{(2+\eta(\lambda_1+\lambda_2))\Delta}\right]-1}{\delta}}\Big]^2} \\ & + \frac{p_t^2(\lambda_2)(1-|\mathbf{u}_1^T\mathbf{w}_0|^2)}{\delta p_t^2(\lambda_1)\big[|\mathbf{u}_1^T\mathbf{w}_0|-\sqrt{\frac{\exp\left[\frac{\sigma^2\eta t}{(2+\eta(\lambda_1+\lambda_2))\Delta}\right]-1}{\delta}}\big]^2}.\end{aligned} \quad (5.44)$$

If we have $\eta\sigma^2<\frac{(2+\eta(\lambda_1+\lambda_2))\Delta}{t}$, then we have $\exp\left(\frac{\sigma^2\eta t}{(2+\eta(\lambda_1+\lambda_2))\Delta}\right)-1\leq\frac{2\sigma^2\eta t}{(2+\eta(\lambda_1+\lambda_2))\Delta}$ and also $\exp\left(\frac{\sigma^2\eta t}{(2+\eta(\lambda_1+\lambda_2))\Delta}\right)-1\geq\frac{\sigma^2\eta t}{(2+\eta(\lambda_1+\lambda_2))\Delta}$. Therefore, with probability at least $1-2\delta$,

$$\begin{aligned}1-\frac{\langle\mathbf{u}_1,\mathbf{w}_t'\rangle^2}{\|\mathbf{w}_t'\|^2} \leq & \frac{\sqrt{d}2\sigma^2\eta t}{\delta(2+\eta(\lambda_1+\lambda_2))\Delta\left[|\mathbf{u}_1^T\mathbf{w}_0|-\sqrt{\frac{\sigma^2\eta t}{\delta(2+\eta(\lambda_1+\lambda_2))\Delta}}\right]^2}+ \\ & \frac{(1-|\mathbf{u}_1^T\mathbf{w}_0|^2)}{\delta\left[|\mathbf{u}_1^T\mathbf{w}_0|-\sqrt{\frac{\sigma^2\eta t}{\delta(2+\eta(\lambda_1+\lambda_2))\Delta}}\right]^2}\frac{p_t^2(\lambda_2)}{p_t^2(\lambda_1)}.\end{aligned} \quad (5.45)$$

Furthermore, if we have $\eta\sigma^2<\frac{(2+\eta(\lambda_1+\lambda_2))\cdot\Delta\cdot(|\mathbf{u}_1^T\mathbf{w}_0|-\sqrt{\iota})^2}{t}$ for some $\iota>0$, then we have

$$\left[|\mathbf{u}_1^T\mathbf{w}_0|-\sqrt{\frac{\sigma^2\eta t}{\delta(2+\eta(\lambda_1+\lambda_2))\Delta}}\right]^2>\iota. \quad (5.46)$$

Therefore, we have

$$1-\frac{\langle\mathbf{u}_1,\mathbf{w}_t'\rangle^2}{\|\mathbf{w}_t'\|^2}\leq\frac{\sqrt{d}2\sigma^2\eta t}{\delta(2+\eta(\lambda_1+\lambda_2))\Delta\iota}+\frac{(1-|\mathbf{u}_1^T\mathbf{w}_0|^2)}{\delta\iota}\frac{p_t^2(\lambda_2)}{p_t^2(\lambda_1)}. \quad (5.47)$$

plugging the bound on $\frac{p_t^2(\lambda_2)}{p_t^2(\lambda_1)}$ in Lemma 20 into (5.47) results in the bound in the statement of Theorem 9. □

Next, we show that if we know the iteration budget ahead of time, we can choose a fixed stepsize such that the HBOR algorithm obtains near optimal (up to polylog term) decay in variance term (but non-optimal decay in bias term).

Corollary 2. *Consider the setting of Theorem 9. Given a budget of T iterations, with the choice of stepsize $\eta = (\frac{p\log T}{\sqrt{2\Delta T}})^2 \frac{1}{\lambda_1}$, we have the following error after $T > \frac{2p}{\sqrt{2\Delta}} \log \frac{p}{\sqrt{2\Delta}}$ iterations:*

$$e_T \leq \frac{9\sqrt{d}\sigma^2}{2\Delta^2\delta\iota}\frac{\log^2 T}{T} + \frac{e_0}{\delta\iota T^{p'-2}}, \tag{5.48}$$

where $p' = \frac{p}{\sqrt{\sqrt{\lambda_1}}}$.

Proof. It follows from Theorem 9 that

$$e_t \leq K_1\eta T + K_2\mathbf{e}_0 T^2 e^{-\gamma\sqrt{\eta}T}. \tag{5.49}$$

It follows from $T > \frac{2p}{\sqrt{2\Delta}} \log \frac{p}{\sqrt{2\Delta}}$ that $\frac{p\log T}{\sqrt{2\Delta T}} \leq 1$. Furthermore, we have $\gamma \leq \sqrt{2\Delta}$. Then, we have

$$\begin{aligned} e_t &\leq K_1\eta T + K_2\mathbf{e}_0 T^2 e^{-\frac{p}{\sqrt{\lambda_1}}\log T} \\ &\leq K_1\frac{p^2\log^2 T}{2\Delta\cdot T}\frac{1}{\lambda_1} + K_2 e_0 T^{2-\frac{p}{\sqrt{\lambda_1}}}. \end{aligned}$$

By setting $p = p'\sqrt{\lambda_1}$, we have

$$e_T \leq \frac{9\sqrt{d}\sigma^2}{2\Delta^2\delta\iota}\frac{\log^2 T}{T} + \frac{e_0}{\delta\iota T^{p'-2}}. \tag{5.50}$$

□

5.4 Multistage HB Accelerated PCA

We showed in the last section that accelerated PCA with $O(\frac{\log^2 T}{T^2})$ step-size results in an extra $(\frac{\log^2 T}{T^2})$ rate for the bias term and $(\frac{\log^2 T}{T^2})$. In this section, we follow the method proposed in Aybat et al. [155] to break the optimization process into multiple stages

in hope of improving the convergence rate of accelerated PCA. This method, which we call Multistage heavy ball-accelerated Oja's rule (MHBOR), consists of successive runs of HBOR with (different) fixed stepsizes. Specifically, the estimates are generated by the following rule.

$$\mathbf{w}'_{t_k} = (\mathbf{I} + \eta_k \mathbf{A}_{t_k})\mathbf{w}_{t_k-1} - \beta_k \mathbf{w}_{t_k-2}, \qquad \mathbf{w}_{t_k} = \mathbf{w}'_{t_k}/\|\mathbf{w}'_{t_k}\|.$$

where $1 \leq t_k \leq T_k$ with

$$\begin{aligned} k = 1: \qquad & \eta_1 = \bar{\eta} \leq \frac{1}{\lambda_1}, \qquad T_1 \geq 1, \\ k > 1: \qquad & \eta_k = \bar{\eta}/2^{2k}, \qquad T_k = (1+c)2^k \bar{T} \triangleq (1+c)2^k \lceil \frac{1}{\gamma_k \sqrt{\bar{\eta}}} \log(K_2 2^p) \rceil. \end{aligned} \tag{5.51}$$

and $\beta_k = (1+\eta_k\lambda_2)^2/4$ and $c > 0$.

Remark. The learning rate and the number of iterations are inspired by Aybat et al. [155]. The $(1+c)^2$ term is intended to overcome the difficulty caused by the $(T+1)^2/\sum_{t=0}^{T} a^{-2t}$ term on the RHS of (5.42).

Before stating the convergence result of MHBOR, we present Lemma 21 that will prove useful in establishing the convergence result. We know from Theorem 9 that in the k-th stage of MHBOR, we have

$$\mathbf{e}_T^k \leq K_1 \eta T + K_2 \mathbf{e}_0^k \Big(\frac{T+1}{\sum_{t=0}^{T} a^{-2t}} \Big)^2 e^{-\gamma\sqrt{\bar{\eta}}t}, \tag{5.52}$$

where $\mathbf{e}_0^k$ is the error in the beginning of the the k-th stage and $\mathbf{e}_t^k$ is the the error after t iterations in the k-th stage. In Lemma 21, we study how the term $\frac{(T+1)^2}{\sum_{t=0}^{T} a^{-2t}}$ looks in this scheme.

Lemma 21. *Consider a PSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ with eigenvalues $1 \geq \lambda_1 \geq \cdots \geq \lambda_d$ and eigengap $\Delta \triangleq \lambda_1 - \lambda_2$. Assume that estimates $\mathbf{A}_t$ of matrix $\mathbf{A}$ satisfy the set of assumptions in (5.4). Consider the estimates stepsize η_k and epoch length T_k as described in (5.51). Suppose that we have $\bar{\eta}\sigma^2 < \frac{(2+\bar{\eta}(\lambda_1+\lambda_2))\cdot\Delta\cdot(|\mathbf{u}_1^T\mathbf{w}_0|-\sqrt{\iota})^2}{T_1}$ for some*

$\iota > 0$. *Define* $a_k \triangleq \frac{(1+\eta_k\lambda_1)+\sqrt{(1+\eta_k\lambda_1)^2-(1+\eta_k\lambda_2)^2}}{1+\eta_k\lambda_2}$. *Then, we have*

$$\frac{T_k}{\sum_{t=0}^{T_k-1} a_k^{-2t}} \leq 16\bar{T}. \tag{5.53}$$

Proof of Lemma 21. In the multistage method, the values of η, and consequently the values of μ_2, μ_{1_+} and a, change at every stage. At stage k, the value we have

$$\begin{aligned} a_k = \frac{\mu_{1_+,k}}{\mu_{2,k}} &= \frac{(1+\eta_k\lambda_1)+\sqrt{(1+\eta_k\lambda_1)^2-(1+\eta_k\lambda_2)^2}}{1+\eta_k\lambda_2} \\ &= \frac{1+\bar{\eta}\lambda_1/2^{2k}+\sqrt{\bar{\eta}\Delta}\sqrt{2+\frac{\bar{\eta}}{2^{2k}}(\lambda_1+\lambda_2)}/2^k}{1+\bar{\eta}\lambda_2/2^{2k}}. \end{aligned} \tag{5.54}$$

Let us find a lower bound on

$$\sum_{n=0}^{t} a_k^{-2n} = \frac{1-a_k^{-2(t+1)}}{1-a_k^{-2}}.$$

Define $\tau \triangleq 2^{2k}$ for $k > 1$. We have

$$\begin{aligned} a_k^{-2} &= \left(\frac{1+\bar{\eta}\lambda_1/\tau^2+\sqrt{\bar{\eta}\Delta}\sqrt{2+\frac{\bar{\eta}}{\tau^2}(\lambda_1+\lambda_2)}/\tau}{1+\bar{\eta}\lambda_2/\tau^2}\right)^{-2} \\ &= \left(1+\frac{\bar{\eta}\Delta/\tau^2+\sqrt{\bar{\eta}\Delta}\sqrt{2+\frac{\bar{\eta}}{\tau^2}(\lambda_1+\lambda_2)}/\tau}{1+\bar{\eta}\lambda_2/\tau^2}\right)^{-2} \\ &\geq \left(1+\frac{\bar{\eta}\Delta/\tau^2+2\sqrt{\bar{\eta}\Delta}/\tau}{1+\bar{\eta}\lambda_2/\tau^2}\right)^{-2} \\ &\geq \left(1+\frac{3\sqrt{\bar{\eta}\Delta}/\tau}{1+\bar{\eta}\lambda_2/\tau^2}\right)^{-2} \\ &\geq \left(1+\frac{3\sqrt{\bar{\eta}\Delta}}{\tau}\right)^{-2}. \end{aligned} \tag{5.55}$$

Note that for $K > 1$, the final bound can be improved to $\left(1 + \frac{2\sqrt{\bar{\eta}\bar{\Delta}}}{\tau}\right)^{-2}$. Moreover,

$$\begin{aligned}
a_k^{-2T_k} &= \left(\frac{1 + \bar{\eta}\lambda_1/\tau^2 + \sqrt{\bar{\eta}\bar{\Delta}}\sqrt{2 + \frac{\bar{\eta}}{\tau^2}(\lambda_1 + \lambda_2)}/\tau}{1 + \bar{\eta}\lambda_2/\tau^2}\right)^{-2T_k} \\
&= \left(1 + \frac{\bar{\eta}\Delta/\tau^2 + \sqrt{\bar{\eta}\bar{\Delta}}\sqrt{2 + \frac{\bar{\eta}}{\tau^2}(\lambda_1 + \lambda_2)}/\tau}{1 + \bar{\eta}\lambda_2/\tau^2}\right)^{-2T_k} \\
&\leq \left(1 + \frac{\bar{\eta}\Delta/\tau^2 + \sqrt{2\bar{\eta}\bar{\Delta}}/\tau}{2}\right)^{-2T_k} \\
&= \frac{1}{\left(1 + \frac{\bar{\eta}\Delta/\tau^2 + \sqrt{2\bar{\eta}\bar{\Delta}}/\tau}{2}\right)^{2T_k}} \\
&\overset{(a)}{\leq} \frac{1}{1 + \bar{\eta}\Delta T_k/\tau^2 + \sqrt{2\bar{\eta}\bar{\Delta}}T_k/\tau},
\end{aligned} \tag{5.56}$$

where (a) is due to Bernoulli's inequality. Define $u \triangleq \sqrt{\bar{\eta}\bar{\Delta}}$. For $K = 1$ we have

$$\begin{aligned}
\sum_{n=0}^{T_1} a_1^{-2n} &= \frac{1 - a_1^{-2(T_1+1)}}{1 - a_1^{-2}} \\
&\geq \frac{1 - \frac{1}{1 + \bar{\eta}\Delta T_1 + \sqrt{2\bar{\eta}\bar{\Delta}}T_1}}{1 - \left(1 + 3\sqrt{\bar{\eta}\bar{\Delta}}\right)^{-2}} \\
&= \frac{\frac{u^2T_1 + \sqrt{2}uT_1}{1 + u^2T_1 + \sqrt{2}uT_1}}{1 - \frac{1}{(1+3u)^2}} \\
&= \frac{\left(u^2 + \sqrt{2}u\right)T_1(1+3u)^2}{\left(1 + u^2T_1 + \sqrt{2}uT_1\right)(6u + 9u^2)} \\
&\geq \frac{T_1(1+3u)^2}{9\left(1 + u^2T_1 + \sqrt{2}uT_1\right)} \\
&\geq \frac{(1+3u)^2}{9\left(1 + u^2 + \sqrt{2}u\right)} \\
&\geq \frac{(1+3u)^2}{4\left(1+u\right)^2} \\
&\geq \frac{1}{4},
\end{aligned} \tag{5.57}$$

which is trivial since we know $\sum_{n=0}^{T_1} a_1^{-2n} > 1$. However, for $K > 1$ the bound is

nontrivial:

$$\begin{aligned}
\sum_{n=0}^{T_k} a_k^{-2n} &= \frac{1-a_k^{-2(T_k+1)}}{1-a_k^{-2}} \\
&\geq \frac{1-\frac{1}{1+\bar{\eta}\Delta(1+c)\bar{T}/\tau+\sqrt{2\bar{\eta}\Delta}(1+c)\bar{T}}}{1-\left(1+\frac{2\sqrt{\bar{\eta}\Delta}}{\tau}\right)^{-2}} \\
&= \frac{(1+c)\frac{u^2\bar{T}/\tau+\sqrt{2}u\bar{T}}{1+u^2\bar{T}/\tau+\sqrt{2}u\bar{T}}}{1-\frac{1}{(1+2u/\tau)^2}} \\
&= \frac{(1+c)\left(u^2/\tau+\sqrt{2}u\right)\bar{T}(1+2u/\tau)^2}{\left(1+u^2\bar{T}/\tau+\sqrt{2}u\bar{T}\right)\left(4u/\tau+4u^2/\tau^2\right)} \\
&\geq \frac{(1+c)\tau\bar{T}(1+2u/\tau)^2}{4\left(1+u^2\bar{T}\tau+\sqrt{2}u\bar{T}\right)} \\
&\geq \frac{(1+c)\tau(1+u/\tau)^2}{4\left(1+u^2/\tau+\sqrt{2}u\right)} \\
&\geq \frac{(1+c)\tau}{4\left(1+u\right)^2}. \qquad (5.58)
\end{aligned}$$

It follows immediately from this that

$$\begin{aligned}
\frac{T_k}{\sum_{n=0}^{T_k-1} a_k^{-2n}} &\leq \frac{(1+c)\tau\bar{T}}{\frac{(1+c)\tau}{4(1+u)^2}} \\
&\leq 4(1+\sqrt{\bar{\eta}\Delta})^2\bar{T} \\
&\overset{(a)}{\leq} 16\bar{T}, \qquad (5.59)
\end{aligned}$$

where (a) follows from $\bar{\eta} \leq 1/\lambda_1$. □

Now, we are ready to state and prove the convergence rate of MHBOR. First, Theorem 10 states the suboptimality error at the end of each stage.

Theorem 10. *Consider a PSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ with eigenvalues $1 \geq \lambda_1 \geq \cdots \geq \lambda_d$ and eigengap $\Delta := \lambda_1 - \lambda_2$. Assume that estimates $\mathbf{A}_t$ of matrix $\mathbf{A}$ satisfy the set of assumptions in* (5.4)*. Consider running the MHBOR algorithm with parameters described in* (5.51) *such that $\bar{\eta}\sigma^2 < \frac{(2+\bar{\eta}(\lambda_1+\lambda_2))\cdot\Delta\cdot(|\mathbf{u}_1^T\mathbf{w}_0|-\sqrt{\iota})^2}{T_1}$ for some $\iota > 0$. Assume that Then, with probability at least $1-2k\delta$, the error at the end of the k-th stage,*

$\mathbf{e}_t^k \triangleq 1 - \frac{\left\langle \mathbf{u}_1, \mathbf{w}_t'^k \right\rangle^2}{\|\mathbf{w}_t'^k\|^2}$, *is*

$$\mathbf{e}_{T_k+1}^k \leq \frac{1}{2^{(k-1)p}} K_2 \mathbf{e}_0 T_1^2 e^{-\gamma\sqrt{\bar{\eta}}T_1} + \frac{1}{2^{(k-1)}} K_1(1+c)\bar{\eta}\bar{T}, \tag{5.60}$$

where $K_1 = \frac{\sqrt{d}\sigma^2}{\delta\Delta\iota}$, $K_2 = \frac{1}{\delta\iota}$, $\gamma = \sqrt{2\Delta}$, *and* c *is chosen sufficiently large such that* $(1+c)^2 16^2 \bar{T}^2 \leq e^c$.

Proof. We use induction to prove this theorem. For the first stage (base case), we have

$$\begin{aligned}
\mathbf{e}_{T_1+1}^1 &\leq K_1 \bar{\eta} T_1 + K_2 \mathbf{e}_0 T_1^2 e^{-\gamma_1\sqrt{\bar{\eta}}T_1} \\
&\leq K_1(1+c)\bar{\eta}T_1 + K_2 \mathbf{e}_0 T_1^2 e^{-\gamma_1\sqrt{\bar{\eta}}T_1}.
\end{aligned} \tag{5.61}$$

Next, we study the following stages (induction step). Note that for $k > 1$, we have

$$\begin{aligned}
\mathbf{e}_{T_k+1}^k &\leq K_1 \eta_k T_k + K_2 \mathbf{e}_1^k T_k^2 e^{-\gamma_k\sqrt{\eta_k}T_k} \\
&\leq \frac{1}{2^k} K_1(1+c)\bar{\eta}\bar{T} + K_2 \mathbf{e}_1^k \cdot 16^2 \bar{T}^2 e^{-(1+c)\log(K_2 2^p)} \\
&\leq \frac{1}{2^k} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^p}\mathbf{e}_1^k (1+c)^2 16^2 \bar{T}^2 e^{-c} \\
&\leq \frac{1}{2^k} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^p}\mathbf{e}_1^k \frac{(1+c)^2 16^2 \bar{T}^2}{e^c}.
\end{aligned} \tag{5.62}$$

For large enough c such that $\frac{(1+c)^2 16^2 \bar{T}^2}{e^c} \leq 1$, we have

$$\mathbf{e}_{T_k+1}^k \leq \frac{1}{2^k} K_1 \bar{\eta}\bar{T} + \frac{1}{2^p}\mathbf{e}_1^k. \tag{5.63}$$

If $\mathbf{e}_{T_{k-1}+1}^{k-1} \leq \frac{1}{2^{(k-2)}} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-2)p}} K_2 \mathbf{e}_0 T_1^2 e^{-\gamma\sqrt{\bar{\eta}}T_1}$, we get

$$\begin{aligned}
\mathbf{e}_{T_k+1}^k &\leq \frac{1}{2^k} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^p}\frac{1}{2^{(k-2)}} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^p}\frac{1}{2^{(k-2)p}} K_2 \mathbf{e}_0 T_1^2 e^{-\gamma\sqrt{\bar{\eta}}T_1} \\
&= \frac{1}{2^k} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-2+p)}} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-1)p}} K_2 \mathbf{e}_0 T_1^2 e^{-\gamma\sqrt{\bar{\eta}}T_1} \\
&\overset{(a)}{\leq} \frac{1}{2^k} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^k} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-1)p}} K_2 \mathbf{e}_0 T_1^2 e^{-\gamma\sqrt{\bar{\eta}}T_1} \\
&\leq \frac{1}{2^{k-1}} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-1)p}} K_2 \mathbf{e}_0 T_1^2 e^{-\gamma\sqrt{\bar{\eta}}T_1} \\
&\leq \frac{1}{2^{(k-1)}} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-1)p}} K_2 \mathbf{e}_0 T_1^2 e^{-\gamma\sqrt{\bar{\eta}}T_1},
\end{aligned} \tag{5.64}$$

where inequality (a) follows from $p \geq 2$. Therefore, we showed inductively that at the end of the k-th stage we have

$$\mathbf{e}_{T_k+1}^k = 1 - \frac{(\mathbf{u}_1^T \mathbf{w}'^k_{T_k+1})^2}{\|\mathbf{w}'^k_{T_k+1}\|^2} \leq \frac{1}{2^{(k-1)}} K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-1)p}} K_2 \mathbf{e}_0 T_1^2 e^{-\gamma\sqrt{\bar{\eta}}T_1}. \quad (5.65)$$

□

Finally, Theorem 11 states an upper bound on the suboptimality error of MHBOR at any iteration t.

Theorem 11. *Consider a PSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ with eigenvalues $1 \geq \lambda_1 \geq \cdots \geq \lambda_d$ and eigengap $\Delta := \lambda_1 - \lambda_2$. Assume that estimates $\mathbf{A}_t$ of matrix $\mathbf{A}$ satisfy the set of assumptions in (5.4). Consider the estimates generated by the multi-stage accelerated Oja's method with heavy-balling (MHBOR)*

$$\mathbf{w}'_{t_k} = (\mathbf{I} + \eta_k \mathbf{A}_{t_k})\mathbf{w}_{t_k-1} - \beta_k \mathbf{w}_{t_k-2}, \qquad \mathbf{w}_{t_k} = \mathbf{w}'_{t_k}/\|\mathbf{w}'_{t_k}\|.$$

where $1 \leq t_k \leq T_k$ with

$$\begin{aligned} k = 1: &\qquad \eta_1 = \bar{\eta} \leq \frac{1}{\lambda_1}, \qquad T_1 \geq 1, \\ k > 1: &\qquad \eta_k = \bar{\eta}/2^{2k}, \qquad T_k = (1+c)2^k \bar{T} \triangleq (1+c)2^k \lceil \frac{1}{\gamma\sqrt{\bar{\eta}}} \log(K_2 2^p) \rceil. \end{aligned} \quad (5.66)$$

and $\beta_k = (1+\eta_k\lambda_2)^2/4$ and $c > 0$. Further, suppose that for some $\iota > 0$ we have $\bar{\eta}\sigma^2 < \frac{(2+\bar{\eta}(\lambda_1+\lambda_2))\cdot\Delta\cdot(|\mathbf{u}_1^T\mathbf{w}_0| - \sqrt{\iota})^2}{T_1}$. Then, after T iterations of the MHBOR algorithm, with probability at least $1 - 2\lceil \log T \rceil \delta$ for some $\delta > 0$, we have

$$\begin{aligned} 1 - \langle \mathbf{u}_1, \mathbf{w}_T \rangle^2 \leq & \frac{10(1+c)^2 K_1 \sqrt{\bar{\eta}} \lceil \frac{1}{\gamma} \log(K_2 2^p) \rceil^2}{T - T_1} \\ & + \frac{64(1+c)^2 \lceil \frac{1}{\gamma\sqrt{\bar{\eta}}} \log(K_2 2^p) \rceil^2}{(T-T_1)^2} K_2 \left(1 - \langle \mathbf{u}_1, \mathbf{w}_0 \rangle^2\right) T_1^2 e^{-\gamma\sqrt{\bar{\eta}}T_1}, \end{aligned} \quad (5.67)$$

where $K_1 = \frac{\sqrt{d}\sigma^2}{\delta\Delta\iota}$, $K_2 = \frac{1}{\delta\iota}$, and $\gamma = \sqrt{2\Delta}$.

Proof of Theorem 11. First, let us find an upper bound on the error $\mathbf{e}_m^k$ using the bound

$\mathbf{e}_t \leq K_1\eta t + K_2\mathbf{e}_0\frac{p_t(\lambda_2)}{p_t(\lambda_1)}$ found in inequality (5.47). We have

$$\begin{aligned}
\mathbf{e}_m^k &\leq K_1\eta_k m + K_2\frac{p_m(\lambda_2)}{p_m(\lambda_1)}\mathbf{e}_0^k \\
&\leq K_1\eta_k m + K_2\frac{p_m(\lambda_2)}{p_m(\lambda_1)}\mathbf{e}_{T_{k-1}}^{k-1} \\
&\overset{(a)}{\leq} K_1\eta_k m + \left(\frac{1}{2^{(k-2)}}K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-2)p}}K_2\mathbf{e}_0T_1^2e^{-\gamma\sqrt{\bar{\eta}}T_1}\right) \\
&\leq K_1\eta_k T_k + \left(\frac{1}{2^{(k-2)}}K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-2)p}}K_2\mathbf{e}_0T_1^2e^{-\gamma\sqrt{\bar{\eta}}T_1}\right) \\
&\leq \frac{1}{2^k}K_1\bar{\eta}\bar{T} + \left(\frac{1}{2^{(k-2)}}K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-2)p}}K_2\mathbf{e}_0T_1^2e^{-\gamma\sqrt{\bar{\eta}}T_1}\right) \\
&\leq \frac{1}{2^{(k-2)}}\frac{5}{4}K_1(1+c)\bar{\eta}\bar{T} + \frac{1}{2^{(k-2)p}}K_2\mathbf{e}_0T_1^2e^{-\gamma\sqrt{\bar{\eta}}T_1},
\end{aligned} \tag{5.68}$$

where (a) follows from Theorem 10. Now, we need to the the corresponding indices k and m such that $\mathbf{w}_m^k$ corresponds to $\mathbf{w}_t$. Remember that $T_k = (1+c)2^k\bar{T}$. It easily follows that for $t > T_1$

$$\begin{aligned}
&\tau_{K+1} \triangleq \sum_{i=1}^{K+1} T_k = T_1 + (1+c)(2^{K+2}-4)\bar{T} \\
\Rightarrow &t - T_1 \leq (1+c)(2^{K+2}-4)\bar{T} \\
\Rightarrow &\frac{1}{2^{K+2}} \leq \frac{1}{2^{K+2}-4} \leq \frac{(1+c)\bar{T}}{t-T_1} \\
\Rightarrow &\frac{1}{2^{K-1}} \leq \frac{8(1+c)\bar{T}}{t-T_1}.
\end{aligned} \tag{5.69}$$

Since $p \geq 2$, it follows from (5.68) that

$$\mathbf{e}_m^{K+1} \leq \frac{1}{2^{(K-1)}}K_1\frac{5(1+c)}{4}\bar{\eta}\bar{T} + \frac{1}{2^{(K-1)p}}K_2\mathbf{e}_0T_1^2e^{-\gamma\sqrt{\bar{\eta}}T_1}. \tag{5.70}$$

Let us set $p = 2$. Plugging (5.69) into (5.65) results in

$$\begin{aligned}
\mathbf{e}_t = \mathbf{e}_{t-\Gamma_K}^{K+1} &\leq \frac{1}{2^{(K-1)}}\frac{5(1+c)}{4}K_1\bar{\eta}\bar{T} + \frac{1}{2^{2(K-1)}}K_2\mathbf{e}_0T_1^2e^{-\gamma\sqrt{\bar{\eta}}T_1} \\
&\leq \frac{10K_1'\bar{\eta}\bar{T}^2}{t-T_1} + \frac{64(1+c)^2\bar{T}^2}{(t-T_1)^2}K_2\mathbf{e}_0T_1^2e^{-\gamma\sqrt{\bar{\eta}}T_1}.
\end{aligned} \tag{5.71}$$

where $K_1' = (1+c)^2 K_1$ and $\Gamma_K = \sum_{k=1}^{K} T_k$. Therefore, for $t > T_1$ we have

$$1 - \langle \mathbf{u}_1, \mathbf{w}_t \rangle^2 \leq \frac{10K_1'\sqrt{\bar{\eta}}\lceil \frac{1}{\gamma}\log(K_2 2^p)\rceil^2}{t - T_1} + \frac{64(1+c)^2\lceil \frac{1}{\gamma\sqrt{\bar{\eta}}}\log(K_2 2^p)\rceil^2}{(t-T_1)^2} K_2 \left(1 - \langle \mathbf{u}_1, \mathbf{w}_0 \rangle^2\right) T_1^2 e^{-\gamma\sqrt{\bar{\eta}}T_1}. \quad (5.72)$$

□

In the following corollary of Theorem 11 we show that our algorithm, up to a logarithmic factor, achieves the error upper bounds $\Omega(\frac{\sigma^2}{(\lambda_1-\lambda_2)^2 t})$ and $\Omega(e_0 e^{-\sqrt{\lambda_1-\lambda_2}t})$ for the noise term and the bias term, respectively.

Corollary 3. *Consider the setting of Theorem 11. Suppose the computational budget of $T = \alpha T_1$ for some $\alpha \geq 2$. Then with probability at least $1 - 2\lceil \log T \rceil \delta$, for $e_T \triangleq 1 - \langle \mathbf{u}_1, \mathbf{w}_T \rangle^2$ we have*

$$e_T \leq C_1 \frac{\sqrt{\sqrt{d}\sigma^2\bar{\eta}}\lceil \log(K_2 2^p)\rceil^2}{\Delta^2 T} + C_2 \lceil \frac{1}{\sqrt{\bar{\eta}}\Delta}\log(K_2 2^p)\rceil^2 K_2 e_0 e^{-c\sqrt{\bar{\eta}}\Delta T}, \quad (5.73)$$

where $K_2 = \frac{1}{\delta\iota}$.

Remark. Note that the probability of the result in Theorem 11 holds with probability at least $1 - 2\lceil \log T \rceil \delta$. In order to boost the probability to $1-\delta$, is to run $O(logT)$ copies of the algorithm, each with $1 - 2\lceil \log T \rceil \delta$ success probability and then output the geometric median of the solutions, which can be done in nearly linear time [177].

5.5 Conclusion and Future Work

In this chapter, we studied the problem of estimating the top eigenvector of the covariance matrix of a multivariate random variable from i.i.d samples in a streaming setting. A well-known and commonly used algorithm for solving this problem is called Oja's method which can be thought of as projected stochastic gradient descent (SGD). Inspired by recent works on accelerating SGD for certain classes of convex problems [154–157], we investigated the effect of applying a momentum-based acceleration method

called the heavy ball method (Polyak momentum) [158, 166] to Oja's method.We proposed a novel accelerated variant of Oja's rule, called MHBOR, that employs a multi-stage scheme for choosing the step size. We showed near-optimal convergence of this multi-stage accelerated algorithm in the true streaming scheme without the need for large mini-batches or variance reduction schemes, a property that distinguishes our algorithm and our analysis from the existing works on accelerating Oja's method.

We prove the convergence of this multi-stage algorithm and show that it approximately (up to a log factor) achieves the $O\big(\frac{\sqrt{d}\sigma^2}{(\lambda_1-\lambda_2)^2 t}\big)$ upper bound in the bias term and $O(e_0 e^{-\sqrt{\lambda_1-\lambda_2}t})$ upper bound in the noise term. When compared to the minimax lower bounds $\Omega(\frac{\sigma^2}{(\lambda_1-\lambda_2)^2 t})$ and $\Omega(e_0 e^{-\sqrt{\lambda_1-\lambda_2}t})$ for the noise term and the bias term respectively, it becomes clear that our bounds for MHBOR are optimal up to a logarithmic factor (as well as a $\sqrt{d}$ factor in the noise term.)

While the dependence of our convergence result on dimensions d is not optimal and there is an extra log factor in dependence on t, our results show that there could be benefit in applying momentum acceleration to stochastic solvers in this structured nonconvex problem.

In terms of future work, improving the analysis of the algorithm to potentially obtain tighter convergence results is a possible direction. Moreover, Aybat et al. [155] show that acceleration can improve robustness to gradient noise power of gradient methods in stochastic settings (quantified in terms of asymptotic expected suboptimality of the iterates), at least for certain classes of strongly convex function. Inspired by this result, studying the robustness of the the accelerated and non-accelerated stochastic PCA algorithms is another interesting future direction. For stochastic algorithms, in addition to convergence rate, the robustness is an important criteria when comparing performances of different algorithms.

Bibliography

[1] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Proc. Int. Conf. Advances Neural inf. Process. Syst.*, 2016, pp. 3873–3881.

[2] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi, "Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach," in *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, Apr. 2017, pp. 65–74.

[3] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, p. 717, Apr. 2009.

[4] R. Sun and Z. Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6535–6579, Nov. 2016.

[5] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proc. Int. Conf. Advances Neural inf. Process. Syst.*, 2017, p. 1233–1242.

[6] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[7] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Proc. Int. Conf. Advances Neural inf. Process. Syst.*, 2013, pp. 2796–2804.

[8] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, no. 2, pp. 349–396, 2003.

[9] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.

[10] G. H. Golub and C. F. Van Loan, "An analysis of the total least squares problem," *SIAM J. Numer. Anal.*, vol. 17, no. 6, pp. 883–893, 1980.

[11] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, May 2011.

[12] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1711–1732, 2013.

[13] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

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

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

[16] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 494–504, 2011.

[17] 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.

[18] N. Cressie and H.-C. Huang, "Classes of nonseparable, spatio-temporal stationary covariance functions," *J. Amer. Statist. Assoc.*, vol. 94, no. 448, pp. 1330–1339, 1999.

[19] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, 2010.

[20] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[21] M. Signoretto, R. Van de Plas, B. De Moor, and J. A. K. Suykens, "Tensor versus matrix completion: A comparison with application to spectral data," *IEEE Signal Process. Lett.*, vol. 18, no. 7, pp. 403–406, 2011.

[22] B. Recht, "A simpler approach to matrix completion," *J. Mach. Learn. Res.*, vol. 12, no. Dec, pp. 3413–3430, 2011.

[23] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.

[24] M. Xu, R. Jin, and Z.-H. Zhou, "Speedup matrix completion with side information: Application to multi-label learning," in *Proc. Int. Conf. Advances Neural inf. Process. Syst.*, 2013, pp. 2301–2309.

[25] K. Zhong, P. Jain, and I. S. Dhillon, "Efficient matrix sensing using rank-1 gaussian measurements," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2015, pp. 3–18.

[26] J. Lu, G. Liang, J. Sun, and J. Bi, "A sparse interactive model for matrix completion with side information," in *Proc. Int. Conf. Advances Neural inf. Process. Syst.*, 2016, pp. 4071–4079.

[27] K.-Y. Chiang, C.-J. Hsieh, and I. S. Dhillon, "Matrix completion with noisy side information," in *Proc. Int. Conf. Advances Neural inf. Process. Syst.*, 2015, pp. 3447–3455.

[28] A. Soni, T. Chevalier, and S. Jain, "Noisy inductive matrix completion under sparse factor models," in *IEEE Int. Symp. Inf. Theory*, June 2017, pp. 2990–2994.

[29] A. Eftekhari, D. Yang, and M. B. Wakin, "Weighted matrix completion and recovery with prior subspace information," *IEEE Trans. Inf. Theory*, vol. 64, no. 6, pp. 4044–4071, 2018.

[30] N. Rao, H. Yu, P. Ravikumar, and I. S. Dhillon, "Collaborative filtering with graph information: Consistency and scalable methods," in *Proc. Int. Conf. Advances Neural inf. Process. Syst.*, 2015, pp. 2107–2115.

[31] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[32] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, Mar. 2011.

[33] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, June 2010.

[34] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. Annu. ACM Symp. Theory Comput.*, 2013, pp. 665–674.

[35] J. 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.

[36] K. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific J. Optimization*, vol. 6, pp. 615–640, 2010.

[37] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points — online stochastic gradient for tensor decomposition," in *Proc. Conf. Learn. Theory*, 2015, pp. 797–842.

[38] K. Kawaguchi, "Deep learning without poor local minima," in *Proc. Int. Conf. Advances Neural inf. Process. Syst.*, 2016, pp. 586–594.

[39] C. Yun, S. Sra, and A. Jadbabaie, "Global optimality conditions for deep neural networks," in *Proc. Int. Conf. Learn. Representations*, 2018.

[40] M. Ghassemi, A. Sarwate, and N. Goela, "Global optimality in inductive matrix completion," in *2018 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2226–2230.

[41] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1724–1732.

[42] W. Cheney and D. R. Kincaid, *Linear Algebra: Theory and Applications*. Jones and Bartlett Publishers, Inc., 2008.

[43] Y. C. Eldar, D. Needell, and Y. Plan, "Unicity conditions for low-rank matrix recovery. arxiv preprint," *arXiv preprint arXiv:1103.5479*, 2011.

[44] Y. Dai and H. Li, "Rank minimization or nuclear-norm minimization: Are we solving the right problem?" in *Proc. Int. Conf. Digit. Image Comput.: Techn. Appl.*, Nov. 2014, pp. 1–8.

[45] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[46] E. J. Candes and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2342–2359, Apr. 2011.

[47] W. A. Sutherland, *Introduction to metric and topological spaces*. Oxford University Press, 2009.

[48] D. Shin, S. Cetintas, K.-C. Lee, and I. S. Dhillon, "Tumblr blog recommendation with boosted inductive matrix completion," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2015, p. 203–212.

[49] P. Rai, "Non-negative inductive matrix completion for discrete dyadic data," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2499–2505.

[50] A. Biswas, M. Kang, D. Kim, and J. Gao, "Robust inductive matrix completion strategy to explore associations between lincrnas and human disease phenotypes," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 6, pp. 2066–2077, 2019.

[51] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene–disease associations," *Bioinformatics*, vol. 30, no. 12, pp. i60–i68, 2014.

[52] Y. Wang and E. Elhamifar, "High rank matrix completion with side information," in *Proc. AAAI Conf. Artif. Intell.*, 2018.

[53] K. Zhong, Z. Song, P. Jain, and I. S. Dhillon, "Nonlinear inductive matrix completion based on one-layer neural networks," *arXiv preprint arXiv:1805.10477*, 2018.

[54] T. Zhou, H. Qian, Z. Shen, C. Zhang, and C. Xu, "Tensor completion with side information: a riemannian manifold approach," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3539–3545.

[55] A. Narita, K. Hayashi, R. Tomioka, and H. Kashima, "Tensor factorization using auxiliary information," *Data Mining Knowl. Discovery*, vol. 25, no. 2, pp. 298–324, 2012.

[56] M. Nimishakavi, B. Mishra, M. Gupta, and P. Talukdar, "Inductive framework for multi-aspect streaming tensor completion with side information," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 307–316.

[57] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.

[58] Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "Sample complexity bounds for dictionary learning from vector- and tensor-valued data," in *Information Theoretic Methods in Data Science*. Cambridge, UK: Cambridge University Press, 2019, ch. 5.

[59] C. F. Van Loan, "The ubiquitous Kronecker product," *J. Comput. Appl. Math.*, vol. 123, no. 1, pp. 85–100, Nov. 2000.

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

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

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

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

[64] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds on dictionary learning for tensor data," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2706–2726, Apr. 2018.

[65] G. Beylkin and M. J. Mohlenkamp, "Numerical operator calculus in higher dimensions," *Proc. Nat. Acad. Sci.*, vol. 99, no. 16, pp. 10 246–10 251, 2002.

[66] 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, Nov. 2013.

[67] M. Ghassemi, Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "Stark: Structured dictionary learning through rank-one tensor recovery," in *Proc. IEEE Int. Workshop Comput. Advances Multi-Sensor Adaptive Process.*, 2017, pp. 1–5.

[68] M. Ghassemi, Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Sample complexity bounds for low-separation-rank dictionary learning," in *Proc. IEEE Int. Symp. Inf. Theory*, 2019, pp. 2294–2298.

[69] M. Ghassemi, Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "Learning mixtures of separable dictionaries for tensor data: Analysis and algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 33–48, 2020.

[70] J. Håstad, "Tensor rank is NP-complete," *J. Algorithms*, vol. 11, no. 4, pp. 644–654, 1990.

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

[72] I. V. Oseledets, "Tensor-train decomposition," *SIAM J. Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011.

[73] A. Novikov, D. Podoprikhin, A. Osokin, and D. P. Vetrov, "Tensorizing neural networks," in *Proc. Advances Neural Inf. Process. Syst.* , 2015, pp. 442–450.

[74] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, 2017.

[75] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Proc. Annu. Conf. Learn. Theory*, ser. JMLR: Workshop and Conf. Proc., vol. 35, 2014, pp. 1–28.

[76] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries," in *Proc. Annu. Conf. Learn. Theory*, vol. 35, no. 1, 2014, pp. 1–15.

[77] R. Gribonval, R. Jenatton, and F. Bach, "Sparse and spurious: Dictionary learning with noise and outliers," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6298–6319, Nov. 2015.

[78] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *Appl. Comput. Harmon. Anal.*, vol. 37, no. 3, pp. 464–491, Nov. 2014.

[79] C. F. Caiafa and A. Cichocki, "Multidimensional compressed sensing and their applications," *Data Mining Knowl. Discovery*, vol. 3, no. 6, pp. 355–380, 2013.

[80] E. Schwab, B. D. Haeffele, R. Vidal, and N. Charon, "Global optimality in separable dictionary learning with applications to the analysis of diffusion MRI," *SIAM J. Imag. Sci.*, vol. 12, no. 4, pp. 1967–2008, 2019.

[81] M. Ghassemi, Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "STARK: Structured dictionary learning through rank-one tensor recovery," in *Proc. IEEE Int. Workshop Comput. Advances Multi-Sensor Adaptive Process.*, 2017, pp. 1–5.

[82] 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 Anal. Signal Separation*, 2018, pp. 456–466.

[83] K. Skretting and K. Engan, "Recursive least squares dictionary learning algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2121–2130, 2010.

[84] E. Dohmatob, A. Mensch, G. Varoquaux, and B. Thirion, "Learning brain regions via large-scale online structured sparse dictionary learning," in *Proc. Advances Neural Inf. Process. Syst.*, 2016, pp. 4610–4618.

[85] W. Rudin, *Principles of mathematical analysis*. McGraw-hill New York, 1964, vol. 3.

[86] 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.

[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] K. Wimalawarne, M. Sugiyama, and R. Tomioka, "Multitask learning meets tensor factorization: Task imputation via convex optimization," in *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, 2014, pp. 2825–2833.

[89] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[90] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc.: Ser. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[91] 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, Jan. 2011.

[92] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear multitask learning," in *Proc. Int. Conf. Mach. Learn.*, vol. 28, no. 3, 2013, pp. 1444–1452.

[93] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, p. 025010, Jan. 2011.

[94] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[95] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific Belmont, 1999.

[96] R. J. Tibshirani, "The lasso problem and uniqueness," *Electron. J. Statist.*, vol. 7, pp. 1456–1490, 2013.

[97] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice Hall Englewood Cliffs, NJ, 1989, vol. 23.

[98] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, vol. 1, Nov. 1993, pp. 40–44.

[99] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[100] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. IEEE Int. Conf. Pattern recognit.*, 2010, pp. 2366–2369.

[101] N. Chatterji and P. L. Bartlett, "Alternating minimization for dictionary learning with random initialization," in *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, 2017, pp. 1997–2006.

[102] K. Schnass, "Dictionary learning-from local towards global and adaptive," *arXiv preprint arXiv:1804.07101*, 2018.

[103] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere," in *Proc. IEEE Int. Conf. Sampling Theory Appl.*, 2015, pp. 407–410.

[104] E. Frolov and I. Oseledets, "Tensor methods and recommender systems," *Data Mining Knowl. Discovery*, vol. 7, no. 3, p. e1201, 2017.

[105] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver, "Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering," in *Proc. ACM Conf. Recommender Syst.*, 2010, pp. 79–86.

[106] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with bayesian probabilistic tensor factorization," in *Proc. SIAM Int. Conf. Data Mining*, 2010, pp. 211–222.

[107] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2773–2832, 2014.

[108] A. Anandkumar, D. J. Hsu, M. Janzamin, and S. M. Kakade, "When are overcomplete topic models identifiable? Uniqueness of tensor Tucker decompositions with structured sparsity," in *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, 2013, pp. 1986–1994.

[109] N. Cohen, O. Sharir, and A. Shashua, "On the expressive power of deep learning: A tensor analysis," in *Proc. Conf. Learn. Theory*, 2016, pp. 698–728.

[110] M. Janzamin, H. Sedghi, and A. Anandkumar, "Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods," *arXiv preprint arXiv:1506.08473*, 2015.

[111] V. Tresp, C. Esteban, Y. Yang, S. Baier, and D. Krompaß, "Learning with memory embeddings," *arXiv preprint arXiv:1511.07972*, 2015.

[112] B. Hutchinson, L. Deng, and D. Yu, "Tensor deep stacking networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1944–1957, 2013.

[113] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642.

[114] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "A survey of multilinear subspace learning for tensor data," *Pattern Recognit.*, vol. 44, no. 7, pp. 1540–1551, 2011.

[115] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang, "Multilinear discriminant analysis for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 212–220, 2007.

[116] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1193–1207, 2010.

[117] T. Barker, T. Virtanen, T. Barker, and T. Virtanen, "Blind separation of audio mixtures through nonnegative tensor factorization of modulation spectrograms," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 12, pp. 2377–2389, Dec. 2016.

[118] M. Imaizumi and K. Hayashi, "Doubly decomposing nonparametric tensor regression," in *Proc. Int. Conf. Mach. Learn.*, vol. 48, June 2016, pp. 727–736.

[119] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *J. Amer. Statist. Assoc.*, vol. 108, no. 502, pp. 540–552, 2013.

[120] X. Li, D. Xu, H. Zhou, and L. Li, "Tucker tensor regression and neuroimaging analysis," *Statist.Biosci.*, vol. 10, no. 3, pp. 520–545, 2018.

[121] J. Duchi, "Lecture notes for statistics 311/electrical engineering 377," vol. 2, p. 23, 2016.

[122] Q. Zhao, G. Zhou, S. Xie, L. Zhang, and A. Cichocki, "Tensor ring decomposition," *arXiv preprint arXiv:1606.05535*, 2016.

[123] G. Rabusseau and H. Kadri, "Low-rank regression with tensor responses," in *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, 2016, pp. 1867–1875.

[124] E. F. Lock, "Tensor-on-tensor regression," *J. Comput. Graphical Statist.*, vol. 27, no. 3, pp. 638–647, 2018.

[125] W. W. Sun and L. Li, "Store: Sparse tensor response regression and neuroimaging analysis," *J. Mach. Learn. Res.*, vol. 18, no. 1, p. 4908–4944, Jan. 2017.

[126] T. Ahmed, H. Raja, and W. U. Bajwa, "Tensor regression using low-rank and sparse tucker decompositions," *SIAM J. Math. Data Sci.*, vol. 2, no. 4, pp. 944–966, 2020.

[127] L. He, K. Chen, W. Xu, J. Zhou, and F. Wang, "Boosted sparse and low-rank tensor regression," in *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, 2018, pp. 1009–1018.

[128] R. Yu and Y. Liu, "Learning from multiway data: Simple and efficient tensor regression," in *Proc. Int. Conf. Mach. Learn.*, vol. 48, June 2016, pp. 373–381.

[129] G. Raskutti and M. Yuan, "Convex regularization for high-dimensional tensor regression," *arXiv preprint arXiv:1512.01215*, vol. 639, 2015.

[130] H. Chen, G. Raskutti, and M. Yuan, "Non-convex projected gradient descent for generalized low-rank tensor regression," *J. Mach. Learn. Res.*, vol. 20, no. 1, p. 172–208, Jan. 2019.

[131] R. Guhaniyogi, S. Qamar, and D. B. Dunson, "Bayesian tensor regression," *J. Mach. Learn. Res.*, vol. 18, no. 1, p. 2733–2763, Jan. 2017.

[132] I. Perros, R. Chen, R. Vuduc, and J. Sun, "Sparse hierarchical tucker factorization and its application to healthcare," in *IEEE Int. Conf. Data Mining*, 2015, pp. 943–948.

[133] B. Hao, B. Wang, P. Wang, J. Zhang, J. Yang, and W. W. Sun, "Sparse tensor additive regression," *arXiv preprint arXiv:1904.00479*, 2019.

[134] K. Wimalawarne, R. Tomioka, and M. Sugiyama, "Theoretical and experimental analyses of tensor-based regression and classification," *Neural Comput.*, vol. 28, no. 4, pp. 686–715, 2016.

[135] T. Suzuki, "Convergence rate of bayesian tensor estimator and its minimax optimality," in *Proc. Int. Conf. Mach. Learn.*, vol. 37, July 2015, p. 1273–1282.

[136] R. Z. Khas'minskii, "A lower bound on the risks of non-parametric estimates of densities in the uniform metric," *Theory Probability Appl.*, vol. 23, no. 4, pp. 794–798, 1979.

[137] B. Yu, "Assouad, fano, and le cam," in *Festschrift for Lucien Le Cam*. Springer, 1997, pp. 423–435.

[138] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.

[139] 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, 2016.

[140] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds for Kronecker-structured dictionary learning," in *Proc. IEEE Int. Symp. Inf. Theory*, July 2016, pp. 1148–1152.

[141] G. Lugosi, "Concentration-of-measure inequalities," *Lecture Notes [Online]*, 2004.

[142] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[143] 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.

[144] J.-L. Durrieu, J.-P. Thiran, and F. Kelly, "Lower and upper bounds for approximation of the kullback-leibler divergence between gaussian mixture models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 4833–4836.

[145] E. Oja, "Simplified neuron model as a principal component analyzer," *J. math. biol.*, vol. 15, no. 3, pp. 267–273, 1982.

[146] E. Oja and J. Karhunen, "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix," *J. math. anal. appl.*, vol. 106, no. 1, pp. 69–84, 1985.

[147] T. Krasulina, "The method of stochastic approximation for the determination of the least eigenvalue of a symmetrical matrix," *USSR Comput. Math. Math. Phys.*, vol. 9, no. 6, pp. 189–195, 1969.

[148] P. Xu, B. He, C. De Sa, I. Mitliagkas, and C. Re, "Accelerated stochastic power iteration," in *Proc. Int. Conf. Artif. Intellig. Statist.*, vol. 84, Apr. 2018, pp. 58–67.

[149] V. Q. Vu, J. Lei *et al.*, "Minimax sparse principal subspace estimation in high dimensions," *Ann. Statist.*, vol. 41, no. 6, pp. 2905–2947, 2013.

[150] P. Jain, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford, "Streaming PCA: Matching matrix bernstein and near-optimal finite sample guarantees for Oja's algorithm," in *Annu. Conf. Learn. Theory*, vol. 49, 23–26 Jun 2016, pp. 1147–1164.

[151] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Netw.*, vol. 2, no. 1, pp. 53 – 58, 1989.

[152] A. Eftekhari and R. A. Hauser, "Principal component analysis by optimization of symmetric functions has no spurious local optima," *SIAM J. Optim.*, vol. 30, no. 1, pp. 439–463, 2020.

[153] A. Gonen and S. Shalev-Shwartz, "Fast rates for empirical risk minimization of strict saddle problems," in *Proc. Conf. Learn. Theory*, vol. 65, July 2017, pp. 1043–1063.

[154] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford, "Accelerating stochastic gradient descent for least squares regression," in *Proc. Conf. Learn. Theory*, vol. 75, July 2018, pp. 545–604.

[155] N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar, "A universally optimal multistage accelerated stochastic gradient method," *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, vol. 32, pp. 8525–8536, 2019.

[156] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 8194–8244, 2017.

[157] N. Loizou and P. Richtárik, "Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods," *Computational Optimization and Applications*, vol. 77, no. 3, pp. 653–710, 2020.

[158] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. Math. Phys.*, vol. 4, no. 5, pp. 1 – 17, 1964.

[159] C. J. Li, M. Wang, H. Liu, and T. Zhang, "Near-optimal stochastic approximation for online principal component estimation," *Math. Program.*, vol. 167, no. 1, pp. 75–97, 2018.

[160] O. Shamir, "Convergence of stochastic gradient descent for PCA," in *Proc. Int. Conf. Mach. Learn.*, vol. 48, June 2016, pp. 257–265.

[161] A. Balsubramani, S. Dasgupta, and Y. Freund, "The fast convergence of incremental PCA," in *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, 2013, pp. 3174–3182.

[162] Z. Allen-Zhu and Y. Li, "First efficient convergence for streaming k-PCA: A global, gap-free, and near-optimal rate," in *Proc. IEEE Annu. Symp. Found. Comput. Sci.*, 2017, pp. 487–492.

[163] O. Shamir, "A stochastic PCA and SVD algorithm with an exponential convergence rate," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, ser. ICML'15, vol. 37, 2015, p. 144–152.

[164] C. Kim and D. Klabjan, "Stochastic variance-reduced heavy ball power iteration," *arXiv preprint arXiv:1901.08179*, 2019.

[165] D. Garber, E. Hazan, C. Jin, Sham, C. Musco, P. Netrapalli, and A. Sidford, "Faster eigenvector computation via shift-and-invert preconditioning," in *Proc. Int. Conf. Mach. Learn.*, vol. 48, June 2016, pp. 2626–2634.

[166] B. T. Polyak, *Introduction to optimization (Translations series in mathematics and engineering)*. Springer, 1987, vol. 1.

[167] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed. Springer, 2014.

[168] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, vol. 28, no. 3, 2013, p. 1139–1147.

[169] C. Hu, W. Pan, and J. T. Kwok, "Accelerated gradient methods for stochastic optimization and online learning," in *Proc. Int. Conf. Advances Neural Inf. Process. Syst.*, 2009, pp. 781–789.

[170] B. Can, M. Gurbuzbalaban, and L. Zhu, "Accelerated linear convergence of stochastic momentum methods in Wasserstein distances," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, June 2019, pp. 891–901.

[171] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Math. Program.*, vol. 156, no. 1–2, p. 59–99, Mar. 2016.

[172] G. Lan and Y. Yang, "Accelerated stochastic algorithms for nonconvex finite-sum and multiblock optimization," *SIAM J. Optim.*, vol. 29, no. 4, pp. 2753–2784, 2019.

[173] Z. Allen-Zhu, "Katyusha X: Simple momentum method for stochastic sum-of-nonconvex optimization," in *Proc. Int. Conf. Mach. Learn.*, vol. 80, July 2018, pp. 179–185.

[174] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson, "Global convergence of the heavy-ball method for convex optimization," in *Proc. Eur. Control Conf.*, 2015, pp. 310–315.

[175] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM J. Optim.*, vol. 26, no. 1, pp. 57–95, 2016.

[176] R. Bardenet and O.-A. Maillard, "Concentration inequalities for sampling without replacement," *Bernoulli*, vol. 21, no. 3, pp. 1361–1385, 2015.

[177] M. B. Cohen, Y. T. Lee, G. Miller, J. Pachocki, and A. Sidford, "Geometric median in nearly linear time," in *Proc. ACM symp. Theory Comput.*, 2016, p. 9–21.

Chapter 6

Appendix

6.1 The Rearrangement Procedure

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 $\mathbf{d}_i = \text{vec}(\mathbf{A}_i)$ for $i = 1, 2$ [59]. Figure 6.1 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[1] 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 Nth 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, starting with the more accessible case of KS matrices that are Kronecker product of $N = 3$ factor matrices and then extending it to the general case. Throughout this section, we define an n-th order "tile" to be a scaled copy of $\mathbf{A}_{N-n+1} \otimes \cdots \otimes \mathbf{A}_N$ for $N > 0$. A zeroth order tile is just an element of a matrix. Moreover, we generalize the concept of slices of a 3rd-order tensor to "hyper-slices": an n-th order hyper-slice is a scaled copy of $\mathbf{d}_N \circ \mathbf{d}_{N-1} \circ \cdots \circ \mathbf{d}_{N-n+1}$

[1] 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.

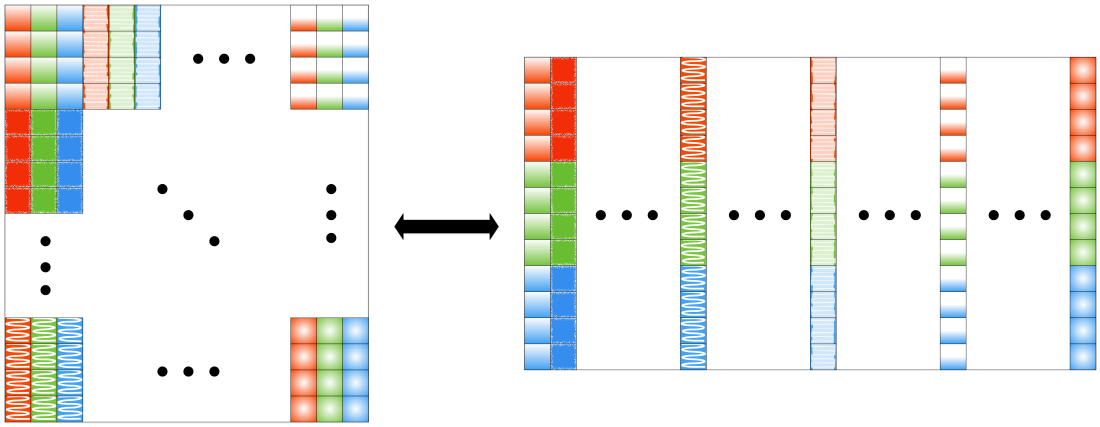


Figure 6.1: Rearranging a Kronecker structured matrix ($N = 2$) into a rank-1 matrix.

6.1.1 Kronecker Product of 3 Matrices

In the case of 3rd-order tensors, we take the following steps:

i) Find index (i, j) in $\mathbf{A}$ that corresponds to the l-th element of $\text{vec}(\mathbf{A})$.

ii) Find the corresponding index (r, c, s) on the third order tensor $\underline{\mathbf{A}}^{\pi}$.

iii) Find the corresponding index l' on $\text{vec}(\underline{\mathbf{A}}^{\pi})$.

iv) Set $\mathbf{\Pi}(l', l) = 1$.

Let $\mathbf{A} = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \mathbf{A}_3$, with $\mathbf{A} \in \mathbb{R}^{m \times p}$ and $\mathbf{A}_i \in \mathbb{R}^{m_i \times p_i}$ for $i \in \{1, 2, 3\}$. For the first operation, we have

$$(i, j) = \left(\left\lceil \frac{l}{m} \right\rceil,\ l - \left\lfloor \frac{l-1}{m} \right\rfloor m \right). \tag{6.1}$$

We can see from Figure 6.2 that 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:

$$T_2 = \underbrace{\left\lfloor \frac{j-1}{p_2 p_3} \right\rfloor}_{S_j^2} m_1 + \underbrace{\left\lfloor \frac{i-1}{m_2 m_3} \right\rfloor}_{S_i^2} + 1, \tag{6.2}$$

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:

$$\begin{aligned} i_2 &= i - S_i^2 m_2 m_3 = i - \left\lfloor \frac{i-1}{m_2 m_3} \right\rfloor m_2 m_3, \\ j_2 &= j - S_j^2 p_2 p_3 = j - \left\lfloor \frac{j-1}{p_2 p_3} \right\rfloor p_2 p_3. \end{aligned} \tag{6.3}$$

For the column index, T_1, we have

$$T_1 = \underbrace{\left\lfloor \frac{j_2-1}{p_3} \right\rfloor}_{S_j^1} m_2 + \underbrace{\left\lfloor \frac{i_2-1}{m_3} \right\rfloor}_{S_i^1} + 1. \tag{6.4}$$

The location of the element on the 1st-order tile is

$$\begin{aligned} i_1 &= i_2 - S_i^1 m_3 = i_2 - \left\lfloor \frac{i_2-1}{m_3} \right\rfloor m_3, \\ j_1 &= j_2 - S_j^1 p_3 = j_2 - \left\lfloor \frac{j_2-1}{p_3} \right\rfloor p_3. \end{aligned} \tag{6.5}$$

Therefore, T_0 can be expressed as

$$T_0 = (j_1 - 1)\, m_3 + i_1. \tag{6.6}$$

Finally, in the last step we find the corresponding index on $\mathrm{vec}(\underline{\mathbf{A}}^{\pi})$ using the

following rule.

$$l' = (T_2 - 1)m_2 m_3 p_2 p_3 + (T_1 - 1)m_3 p_3 + T_0. \tag{6.7}$$

6.1.2 The General Case

We now extend our results to N-th order tensors. Vectorization and its adjoint operation are easy to compute for tensors of any order. We focus on rearranging elements of $\mathbf{A} = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \cdots \otimes \mathbf{A}_N$ to form the N-way rank-1 tensor $\underline{\mathbf{A}}^{\pi}$, where $\mathbf{A}_n \in \mathbb{R}^{m_n \times p_n}$ for $n \in [N]$, $\mathbf{A} \in \mathbb{R}^{m \times p}$, and $\underline{\mathbf{A}}^{\pi} \in \mathbb{R}^{m_N p_N \times m_{N-1} p_{N-1} \times \cdots \times m_1 p_1}$.

We first formally state the rearrangement and then we explain it. Similar to the case of $N = 3$ explained earlier, for each element of the KS matrix $\mathbf{A}$ indexed by (i, j), we first find the $(N-1)$th-order tile to which it belongs, then the $(N-2)$th-order tile, and so on. Let $T_{N-1}, T_{N-2}, \cdots, T_0$ denote the indices of these tiles, respectively. Then, after rearrangement, the element indexed (i, j) on KS matrix $\mathbf{A}$ becomes the element indexed $T_0, \cdots, T_{N-1}$ on the rearrangement tensor $\underline{\mathbf{A}}^{\pi}$.

Now, let us find the indices of the tiles of KS matrix $\mathbf{A}$ to which the element (i, j) belongs. In the following, we denote by (i_n, j_n) the index of this element within its nth-order tile. Note that since $\mathbf{A}$ is an Nth-order tile itself, we can use (i_N, j_N) instead of (i, j) to refer to the index of the element on $\mathbf{A}$ for consistency of notation. For the (i_N, j_N)-th element of $\mathbf{A}$ we have

$$\begin{aligned}
T_{N-1} &= \underbrace{\left\lfloor \frac{j_N - 1}{\Pi_{t=2}^{N} p_t} \right\rfloor}_{S_j^N} m_1 + \underbrace{\left\lfloor \frac{i_N - 1}{\Pi_{t=2}^{N} m_t} \right\rfloor}_{S_i^N} + 1, \\
i_{N-1} &= i_N - S_i^N \, \Pi_{t=2}^{N} m_t. \\
j_{N-1} &= j_N - S_j^N \, \Pi_{t=2}^{N} p_t,
\end{aligned}$$

where T_{N-1} is the index of the $(N-1)$-th order tile and (i_{N-1}, j_{N-1}) is the location of

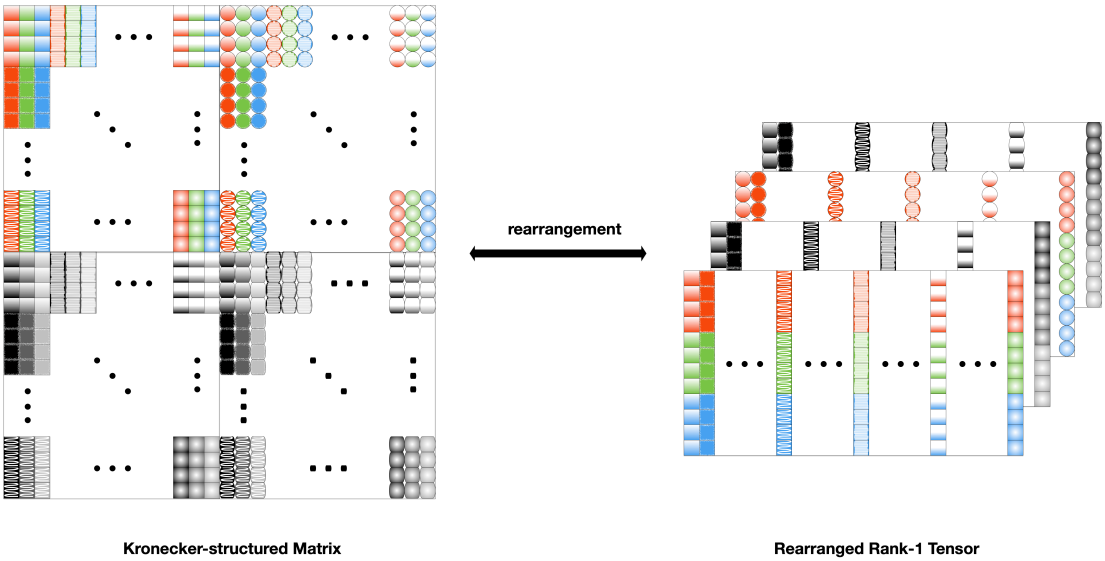


Figure 6.2: Example of rearranging a Kronecker structured matrix ($N = 3$) into a third order rank-1 tensor.

the given element within this tile. Similarly, we have

$$T_{N-n} = \underbrace{\left\lfloor \frac{j_{N-n+1}-1}{\Pi_{t=n+1}^{N}\, p_t} \right\rfloor}_{S_j^{N-n+1}} m_n + \underbrace{\left\lfloor \frac{i_{N-n+1}-1}{\Pi_{t=n+1}^{N}\, m_t} \right\rfloor}_{S_i^{N-n+1}} + 1,$$

$$i_{N-n} = i_{N-n+1} - S_i^n\ \Pi_{t=n+1}^{N}\, m_t,$$

$$j_{N-n} = j_{N-n+1} - S_j^n\ \Pi_{t=n+1}^{N}\, p_t,$$

for $N > n > 1$. Finally, we have

$$T_0 = (j_1 - 1)m_N + i_1.$$

It is now easy to see that the (i_N, j_N)-th element of $\mathbf{A}$ is the $(T_0,\ T_1,\ \cdots,\ T_{N-1})$-th element of $\underline{\mathbf{A}}^{\pi}$.

Intuitively, notice that N-th order KS matrix $\mathbf{A}$ is a tiling of $m_1 \times p_1$ KS tiles of order $N-1$. In rearranging $\mathbf{A}$ into $\underline{\mathbf{A}}^{\pi}$, the elements of each of these $(N-1)$-th order tiles construct a $(N-1)$-th order "hyper-slice". On matrix $\mathbf{A}$, these tiles consist of $m_2 \times p_2$ tiles, each of which is a $(N-2)$th-order KS matrix, whose elements are rearranged to a $(N-2)$-th hyper-slice of $\underline{\mathbf{A}}^{\pi}$, and so on. Hence, the idea is to use the correspondence

between the nth-order tiles and nth-order hyper-slices: finding the index of the n-th order tile of $\mathbf{A}$ on which (i, j) lies is equivalent to finding the index of the nth-order hyper-slice of $\underline{\mathbf{A}}^{\pi}$ to which it is translated. Note that each entry of a tensor in indexed by an N-tuple and the index of an entry of a tensor on its nth hyper-slice is in fact its nth element in the index tuple of this entry. Therefore, we first find the $(N-1)$-th order KS tile of $\mathbf{A}$ on which the (i, j) element lies (equivalent to finding the $(N-1)$th-order hyper-slice to which (i, j) is translated), and then find the location (i_{N-1}, j_{N-1}) of this element on this tile. Next, the $(N-2)$-th order KS tile in which (i_{N-1}, j_{N-1}) lies is found as well as the location (i_{N-2}, j_{N-2}) of the element within this tile, and so on.

6.2 Properties of the Polynomial Sequence

Here, we analyze the polynomial sequence (5.18) in the following lemma (Lemma 22) and its corollary (corollary 4).

Lemma 22. *Given the polynomial sequence $\{p_t(x)\}$ defined as*

$$p_t(x) = (1+\eta x)p_{t-1}(x) - \beta p_{t-2}(x), \quad p_1(x) = 1+\eta x, \quad p_0(x) = 1, \tag{6.8}$$

if $y \neq 4\beta$, we have

$$p_t(x) = \frac{1}{\sqrt{y^2-4\beta}}\left[\left(\frac{y+\sqrt{y^2-4\beta}}{2}\right)^{t+1} - \left(\frac{y-\sqrt{y^2-4\beta}}{2}\right)^{t+1}\right] \tag{6.9}$$

and if $y = 4\beta$,

$$p_t(x) = (t+1)(\sqrt{\beta})^t \tag{6.10}$$

where $y = 1+\eta x$.

Proof of Lemma 22. Consider the generation function of the sequence $\{p_t(x)\}$, i.e.,

$G(x,z) = \sum_{t=0}^{\infty} p_t(x) z^t,\ z \in \mathbb{C}$. It follows from update rule (6.8) that

$$\begin{aligned}
\sum_{t=1}^{\infty} p_{t+1}(x) z^{t+1} &= \sum_{t=1}^{\infty} (1+\eta x) p_t(x) z^{t+1} - \beta \sum_{t=1}^{\infty} p_{t-1}(x) z^{t+1} \\
G(x,z) - p_0 - p_1 z &= (1+\eta x) z (G(x,z) - p_0) - \beta z^2 G(x,z) \\
G(x,z)(1 - (1+\eta x)z + \beta z^2) &= p_0 + (p_1 - (1+\eta x) p_0) z \qquad (6.11)
\end{aligned}$$

Therefore, plugging in the values of p_0 and p_1 results in

$$G(x,z) = \frac{p_0 + (p_1 - (1+\eta x)p_0)z}{1 - (1+\eta x)z + \beta z^2} = \frac{1}{1 - (1+\eta x)z + \beta z^2} \qquad (6.12)$$

Let $y \triangleq 1 + \eta x$. Then,

$$G(y,z) = \frac{1}{1 - yz + \beta z^2} = \frac{1}{\beta (z - r_1)(z - r_2)} \qquad (6.13)$$

where r_1 and r_2 are the roots of $\beta z^2 - yz + 1$, i.e. we have $r_{1,2} = \frac{y \pm \sqrt{y^2 - 4\beta}}{2\beta}$. When

$r_1 \neq r_2$ we have

$$\begin{aligned}
G(y,z) &= \frac{1}{\beta(z-r_1)(z-r_2)} \\
&= \frac{1}{\beta(r_1-r_2)}\left(\frac{1}{r_2-z}-\frac{1}{r_1-z}\right) \\
&= \frac{1}{\beta(r_1-r_2)}\left(\frac{1}{r_2(1-z/r_2)}-\frac{1}{r_1(1-z/r_1)}\right) \\
&\overset{\text{Taylor Expansion}}{=} \frac{1}{\beta(r_1-r_2)}\sum_{t=0}^{\infty}\left(\frac{z^t}{r_2^{t+1}}-\frac{z^t}{r_1^{t+1}}\right) \\
&= \frac{1}{\beta y(r_1-r_2)}\sum_{t=0}^{\infty}\left(\frac{z^t}{r_2^{t+1}}-\frac{z^t}{r_1^{t+1}}\right) \\
&= \frac{1}{\beta(r_1-r_2)}\sum_{t=0}^{\infty}\left(\frac{z^t}{r_2^{t+1}}-\frac{z^t}{r_1^{t+1}}\right) \\
&= \frac{1}{\beta(r_1-r_2)}\sum_{t=0}^{\infty}\left(\frac{z^t}{r_2^{t+1}}-\frac{z^t}{r_1^{t+1}}\right) \\
&= \sum_{t=0}^{\infty}\left[\frac{1/r_2}{\beta(r_1-r_2)}\left(\frac{1}{r_2}\right)^t-\frac{1/r_1}{\beta(r_1-r_2)}\left(\frac{1}{r_1}\right)^t\right]z^t \\
&= \sum_{t=0}^{\infty}\left[\frac{r_1}{r_1-r_2}\left(\frac{1}{r_2}\right)^t-\frac{r_2}{r_1-r_2}\left(\frac{1}{r_1}\right)^t\right]z^t \qquad (6.14)
\end{aligned}$$

Since $r_{1,2} = \frac{y\pm\sqrt{y^2-4\beta}}{2\beta}$, we have $r_1 - r_2 = \frac{\sqrt{y^2-4\beta}}{\beta}$ and $r_1 r_2 = \frac{4\beta}{4\beta^2} = \frac{1}{\beta}$. Therefore, when $|z| < |r_2|$, $G(x,z)$ is well defined, we have

$$\begin{aligned}
p_t(x) &= \frac{r_1}{r_1-r_2}\left(\frac{1}{r_2}\right)^t-\frac{r_2}{r_1-r_2}\left(\frac{1}{r_1}\right)^t \\
&= \frac{r_1}{r_1-r_2}(r_1\beta)^t-\frac{r_2}{r_1-r_2}(r_2\beta)^t \\
&= \frac{r_1}{r_1-r_2}\left(\frac{y+\sqrt{y^2-4\beta}}{2}\right)^t-\frac{r_2}{r_1-r_2}\left(\frac{y-\sqrt{y^2-4\beta}}{2}\right)^t \\
&= \frac{y+\sqrt{y^2-4\beta}}{2\sqrt{y^2-4\beta}}\left(\frac{y+\sqrt{y^2-4\beta}}{2}\right)^t-\frac{y-\sqrt{y^2-4\beta}}{2\sqrt{y^2-4\beta}}\left(\frac{y-\sqrt{y^2-4\beta}}{2}\right)^t \\
&= \frac{1}{\sqrt{y^2-4\beta}}\left[\left(\frac{y+\sqrt{y^2-4\beta}}{2}\right)^{t+1}-\left(\frac{y-\sqrt{y^2-4\beta}}{2}\right)^{t+1}\right]
\end{aligned}$$

On the other hand, when $x =$, we have $y = 2\sqrt{\beta}$ and $r_1 = r_2 = \frac{1}{\sqrt{\beta}}$. Thus,

$$\begin{aligned} G(y,z) =& \frac{1}{\beta(z-r_1)(z-r_2)} \\ =& \frac{1}{(\sqrt{\beta}z-1)^2} \\ \overset{\text{Taylor expansion}}{=}& \sum_{t=0}^{\infty}(t+1)(\sqrt{\beta}z)^t \end{aligned} \tag{6.15}$$

When $z < 1/\sqrt{\beta}$, $G(x,z)$ is well-defined. Therefore, $p_t(\frac{2\sqrt{\beta}-1}{\eta}) = (t+1)(\sqrt{\beta})^t$. □

Corollary 4. *Consider polynomial sequence $\{p_t(x)\}$ defined as in* (6.8)*. Then we have*

$$\prod_{k=1}^{K} p_{t_k}(x) \leq \frac{1}{(\sqrt{y^2-4\beta})^{K-1}} p_{K-1+\sum_{k=1}^{K} t_k}(x) \tag{6.16}$$

Proof of Corrolary 4. We know from (6.9) that

$$\begin{aligned} p_t(x) &= \frac{1}{\beta(r_1-r_2)}(r_1^{t+1} - r_2^{t+1}) \\ &= \frac{1}{\sqrt{y^2-4\beta}} \left[\left(\frac{y+\sqrt{y^2-4\beta}}{2} \right)^{t+1} - \left(\frac{y-\sqrt{y^2-4\beta}}{2} \right)^{t+1} \right] \end{aligned}$$

It follows that

$$\begin{aligned} p_i(x)p_j(x) &= \frac{1}{\beta^2(r_1-r_2)^2}(r_1^{i+1} - r_2^{i+1})(r_1^{j+1} - r_2^{j+1}) \\ &= \frac{1}{\beta^2(r_1-r_2)^2}(r_1^{i+1}r_1^{j+1} + r_2^{i+1}r_2^{j+1} - r_2^{i+1}r_1^{j+1} - r_1^{i+1}r_2^{j+1}) \\ &\overset{(a)}{\leq} \frac{1}{\beta^2(r_1-r_2)^2}(r_1^{i+1}r_1^{j+1} + r_2^{i+1}r_2^{j+1} - r_2^{i+1}r_2^{j+1} - r_2^{i+1}r_2^{j+1}) \\ &= \frac{1}{\beta^2(r_1-r_2)^2}(r_1^{i+j+2} - r_2^{i+j2}) \\ &= \frac{1}{\beta(r_1-r_2)} p_{i+j+1}(x) \\ &= \frac{1}{\sqrt{y^2-4\beta}} p_{i+j+1}(x) \end{aligned} \tag{6.17}$$

Therefore, we can easily use induction to prove the corollary. □