CONVEX AND NONCONVEX OPTIMIZATION GEOMETRIES

by

Qiuwei Li

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Ph.D. (Electrical Engineering).

Golden, Colorado

Date _____

Signed: _____
Qiuwei Li

Signed: _____
Dr. Gongguo Tang
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____
Dr. Peter Aaen
Professor and Department Head
Department of Electrical Engineering

ABSTRACT

Many machine learning and signal processing problems are fundamentally nonconvex. One way to solve them is to transform them into convex optimization problems (a.k.a. convex relaxation), which constitutes a major part of my research. Although the convex relaxation approach is elegant in some ways that it can give information-theoretical sample convexity and minimax denoising rate, but this approach is not efficient in dealing with high-dimensional problems. Therefore, as my second major part of the research, I will directly focus on the fundamentally nonconvex formulations of these nonconvex problems, with a particular interest in understanding the nonconvex optimization landscapes of their fundamental formulations. Then in the third part of my research, I will develop optimization algorithms with provable guarantees that can efficiently navigate these nonconvex landscapes and achieve the global optimality. Finally, in the final part, I will apply the alternating minimization algorithms to general tensor recovery problems and clustering problems.

Part 1: Convex Optimization. In this part, we apply convex relaxations to several popular nonconvex problems in signal processing and machine learning (e.g. line spectral estimation problem and tensor decomposition problem) and prove that the solving the new convex relaxation problems can return the globally optimal solutions of their original nonconvex formulations.

Part 2: Nonconvex Optimization. In this part, we focus on the fundamentally nonconvex optimization landscapes for several low-rank matrix optimization problems with general objective functions, which covers a massive number of popular problems in signal processing and machine learning. In particular, we develop mild conditions for these general low-rank matrix optimization problems to have a benign landscape: all second-order stationary points are global optimal solutions and all saddle points are strict saddles (i.e. Hessian matrix has a negative eigenvalue).

Part 3: Algorithms. In this part, we will develop optimization algorithms with provable second-order optimal convergence for general nonconvex and non-Lipschitz problems. Further, in this part, we also solve an open problem for the second-order convergence of alternating minimization algorithms that have been widely used in practice to solve large-scale nonconvex problems due to their simple implementation, fast convergence, and superb empirical performance. Then the second-order convergence guarantees, along with the knowledge (see Part 2) that a massive number of nonconvex optimization problems have been shown to have a benign landscape (all second-order stationary points are global minima), ensure that the proposed algorithms can find global minima for a class of nonconvex problems.

Part 4: Applications. In this part, we apply the alternating minimization algorithms to several popular applications in signal processing and machine learning, e.g., the low-rank tensor recovery problem and the spherical Principal Component Analysis (PCA).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The best part of graduate school at Mines has been the chance to meet and work with so many amazing people. I have been fortunate to work with a large and talented group of collaborators: Shuang Li, Kai Liu, Ashley Prater-Bennette, Lixin Shen, Youye Xie, Xinshuo Yang, Hua Wang, Zhihui Zhu, and especially Michael Wakin, and Gongguo Tang with whom I collaborate closely on the project—convex and nonconvex optimization geometries (also the title of this thesis).

I wish to express my gratitude to my committee for their valuable suggestions and contributions: to Stephen Pankavich for his enjoyable and inspiring class Math 500 Linear Vector Space; to Tyrone Vincent for his motivating questions and discussions as well as his great course on Estimation theory and Kalman filtering; to Michael Wakin for his helpful discussions and intuitions for my research and writing; and most of all to my advisor Gongguo Tang for his every inspiration in our meetings, his extreme patience in guiding me in the research, and his countless hours devoted to helping me improve my writing and thinking.

I also want to thank all the fun people to work with over the years at Mines: Tong Bai, Armin Eftekhari, Jonathan Helland, Justin Jayne, Shuang Li, Chia Wei Lim, Kai Liu, Weiping Pei, Xinming Wu, Youye Xie, Dehui Yang, Xu Zhou, Zhihui Zhu.

Finally, I want to thank all of my other friends and family for their continued encouragement and support: Grandma, Mom, Dad, Wife, Sister, Gang Li, Yue Wang, Jiaxi Ying, and everyone else who helped me along the way.

CHAPTER 1

INTRODUCTION

This work focuses on using convex and nonconvex optimization methods to model and solve problems in machine learning and signal processing. When we formulate the problem as a convex problem, the statistical performance (cf. 2) can be well analyzed using a suit of powerful convex analysis tools, which have accumulated from several decades of research. For example, a well-designed convex optimization method can achieve information-theoretically optimal sampling complexity, have minimax denoising rate and satisfy tight oracle inequalities. In spite of their optimal statistical performance, the convex optimization methods cannot be scaled to solve the practical problems that originally motivate their development even with specialized first-order algorithms. Further, there are many machine learning and signal processing problems that are fundamentally nonconvex and too expensive/difficult to be convexified. Therefore, as a second part of this work, we focus on the fundamentally nonconvex formulations of some popular machine learning and signal processing problems. In this part, we are particularly interested in understanding the nonconvex optimization landscapes of their fundamental formulations. Then based on this landscape knowledge of these nonconvex optimization problems, in the third part of this work, we focus on developing optimization algorithms with with provable guarantees that can efficiently navigate these nonconvex landscapes and achieve the global optimality. Finally, we some popular applications in signal processing and machine learning are analyzed using the developed optimization algorithms.

Part 1: Convex Optimization

Chapter 2 This chapter investigates the parameter estimation performance of super-resolution line spectral estimation using atomic norm minimization. The focus is on analyzing the algorithm's accuracy of inferring the frequencies and complex magnitudes from noisy observations. When the Signal-to-Noise Ratio is reasonably high and the true frequencies are well separated, we prove that the obtained error bound by the atomic norm estimator matches the Cramér-Rao lower bound up to a logarithmic factor.

Chapter 3 This chapter develops theories and computational methods for guaranteed *overcomplete, non-orthogonal* tensor decomposition using convex optimization. We view tensor decomposition as a problem of measure estimation from moments. We develop a theory for guaranteed decomposition for those tensor factors uniformly distributed on the unit spheres, implying exact decomposition for tensors with random factors. The optimal value of this optimization defines the tensor nuclear norm that can be used to regularize tensor inverse problems, including tensor completion, decisioning, and robust tensor principal component analysis.

Part 2: Nonconvex Optimization

Chapter 4 This chapter considers two popular minimization problems: (i) the minimization of a general convex function $f(\mathbf{X})$ with the domain being positive semi-definite matrices; (ii) the minimization of a general convex function $f(\mathbf{X})$ regularized by the matrix nuclear norm $\|\mathbf{X}\|_*$ with the domain being general matrices. To develop faster and more scalable algorithms, we follow the proposal of Burer and Monteiro to factor the low-rank variable $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$ (for semi-definite matrices) or $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ (for general matrices) and also replace the nuclear norm $\|\mathbf{X}\|_*$ with $(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)/2$. In spite of the non-convexity of the resulting factored formulations, we prove that each critical point either corresponds to the global optimum of the original convex problems or is a strict saddle where the Hessian matrix has a strictly negative eigenvalue.

Chapter 5 This chapter considers the minimization of a general objective function $f(\mathbf{X})$ over the set of rectangular $n \times m$ matrices that have rank at most r. To reduce the computational burden, we factorize the variable $\mathbf{X}$ into a product of two smaller matrices and optimize over these two matrices instead of $\mathbf{X}$. We analyze the global geometry for a general and yet well-conditioned objective function $f(\mathbf{X})$ whose restricted strong convexity and restricted strong smoothness constants are comparable. In particular, we show that the reformulated objective function has no spurious local minima and obeys the *strict saddle property*.

Chapter 6 In this chapter we characterize the global optimization geometry of the nonconvex factored problem and show that the corresponding objective function satisfies the *robust strict saddle property* as long as the original objective function f satisfies restricted strong convexity and smoothness properties, ensuring global convergence of many local search algorithms (such as noisy gradient descent) in polynomial time for solving the factored problem.

Chapter 7 A variety of unconstrained nonconvex optimization problems have been shown to have benign geometric landscapes that satisfy the strict saddle property and have no spurious local minima. We present a general result relating the geometry of an unconstrained centralized problem to its equality-constrained distributed extension. It follows that many global consensus problems inherit the benign geometry of their original centralized counterpart.

Chapter 8 We study the convergence of a variant of distributed gradient descent (DGD) on a distributed low-rank matrix approximation problem wherein some optimization variables are used for consensus (as in classical DGD) and some optimization variables appear only locally at a single node in the network. Using algorithmic connections to gradient descent and geometric connections to the well-behaved landscape of the centralized low-rank matrix approximation problem, we identify sufficient conditions where the new DGD is guaranteed to converge with exact consensus to a global minimizer of the original centralized problem.

For the distributed low-rank matrix approximation problem, these guarantees are stronger—in terms of consensus and optimality—than what appear in the literature for classical DGD and more general problems.

Part 3: Algorithms

Chapter 9 This chapter studies the second-order convergence for both standard alternating minimization and proximal alternating minimization. We show that under mild assumptions on the (nonconvex) objective function, both algorithms avoid strict saddles almost surely from random initialization. Together with known first-order convergence results, this implies both algorithms converge to a second-order stationary point. This solves an open problem for the second-order convergence of alternating minimization algorithms that have been widely used in practice to solve large-scale nonconvex problems due to their simple implementation, fast convergence, and superb empirical performance.

Chapter 10 A crucial and pervasive assumption needed by many modern optimization methods is the global Lipschitz gradient condition. However, many machine learning problems do not admit a globally Lipschitz gradient. In this chapter, we develop and establish second-order convergence guarantees of several Bregman-based methods to deal with general nonconvex objective functions with non-Lipschitz gradients.

Part 4: Applications

Chapter 11 This chapter studies the problem of retrieving a low-rank tensor under a general linear observation model, including both tensor sensing and tensor completion models. Inspired by the superiority of the matrix nuclear norm in low-rank matrix recovery, we will focus on using tensor nuclear norm to regularize the inverse problem of tensor recovery. Unlike the traditional ways of using approximating values of the tensor nuclear norm due to the NP-hardness of computing the tensor nuclear norm, we use the Burer-Monteiro optimization form of the tensor nuclear norm, and we show this form is tight for any randomly generated tensors. Furthermore, we provide an alternating minimization algorithm to solve the tensor nuclear norm regularized problem, as well as the rigorous mathematical analysis of its global convergence.

Chapter 12 Principal Component Analysis (PCA) is one of the most important methods to handle high dimensional data. However, most of the studies on PCA aim to minimize the loss after projection, which usually measure the Euclidean distance, though in some fields, angle distance is known to be more important and critical for analysis. In this chapter, we propose a method by adding constraints on factors to unify the Euclidean distance and angle distance. However, due to the nonconvexity of the objective and constraints, the optimized solution is not easy to obtain. We propose an alternating linearized minimization method to solve it with provable convergence rate and guarantee.

1.1 Related Publications

1. **Q. Li** and G. Tang, “Approximate Support Recovery of Atomic Line Spectral Estimation: A Tale of Resolution and Precision," *IEEE Global Conference on Signal and Information Processing (GlobalSIP 2016)*. [1]

2. **Q. Li** and G. Tang, “Approximate Support Recovery of Atomic Line Spectral Estimation: A Tale of Resolution and Precision," *Applied and Computational Harmonic Analysis*, 2018. [2]

3. **Q. Li**, A. Prater, L. Shen and G. Tang, “Overcomplete Tensor Decomposition via Convex Optimization," *IEEE 6th Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2015)*. [3]

4. **Q. Li**, A. Prater, L. Shen and G. Tang, “A Super-Resolution Framework for Tensor Decomposition," submitted to *Foundations of Computational Mathematics*, 2019. [4]

5. **Q. Li** and G. Tang, “The Nonconvex Geometry of Low-rank Matrix Optimizations with General Objective Functions," *IEEE Global Conference on Signal and Information Processing (GlobalSIP 2017)*. [5]

6. **Q. Li**, Z. Zhu and G. Tang, “The Nonconvex Geometry of Low-rank Matrix Optimization," *Information and Inference: A Journal of the IMA*, 8(1), pp.51-96, 2018. [6]

7. Z. Zhu, **Q. Li**, G. Tang, M. B. Wakin, “Global Optimality in Low-rank Matrix Optimization," *IEEE Global Conference on Signal and Information Processing (GlobalSIP 2017)*. [7]

8. Z. Zhu, **Q. Li**, G. Tang, M. B. Wakin, “Global Optimality in Low-rank Matrix Optimization," *IEEE Transactions on Signal Processing* 66 (13), 3614 - 3628, 2018. [8]

9. Z. Zhu, **Q. Li**, G. Tang and M. B. Wakin, “The Global Optimization Geometry of Nonsymmetric Matrix Factorization and Sensing," under review.

10. **Q. Li**, Z. Zhu, G. Tang, M. B. Wakin, “The Geometry of Equality-Constrained Global Consensus Problems," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*. [9]

11. Z. Zhu,**Q. Li**, X. Yang, G. Tang, M. B. Wakin, “Global Optimality in Distributed Low-rank Matrix Factorization," under review.

12. **Q. Li**, Z. Zhu and G. Tang, “Alternating Minimizations Converge to Second-Order Optimality Solutions," *The 36th International Conference on Machine Learning (ICML 2019)*. [10]

13. **Q. Li**, Z. Zhu, G. Tang, M. B. Wakin, “Provable Bregman-divergence based Methods for Nonconvex and Non-Lipschitz Problems," under review [11].

14. **Q. Li**, K. Liu, G. Tang, H. Wang, “General Tensor Recovery via Alternating Minimization," under review.

15. K. Liu, **Q. Li**, H. Wang, G. Tang, ‘Spherical Principal Component Analysis," *The 2019 SIAM International Conference on Data Mining (SDM19)*. [12]

CHAPTER 2

APPROXIMATE SUPPORT RECOVERY OF ATOMIC LINE SPECTRAL ESTIMATION: A TALE OF RESOLUTION AND PRECISION

This work[1] investigates the parameter estimation performance of super-resolution line spectral estimation using atomic norm minimization. The focus is on analyzing the algorithm's accuracy of inferring the frequencies and complex magnitudes from noisy observations. When the Signal-to-Noise Ratio is reasonably high and the true frequencies are separated by $O(\frac{1}{n})$, the atomic norm estimator is shown to localize the correct number of frequencies, each within a neighborhood of size $O(\sqrt{\log n/n^3}\sigma)$ of one of the true frequencies. Here n is half the number of temporal samples and σ^2 is the Gaussian noise variance. The analysis is based on a primal-dual witness construction procedure. The obtained error bound matches the Cramér-Rao lower bound up to a logarithmic factor. The relationship between resolution (separation of frequencies) and precision or accuracy of the estimator is highlighted. Our analysis also reveals that the atomic norm minimization can be viewed as a convex way to solve a ℓ_1-norm regularized, nonlinear and nonconvex least-squares problem to global optimality.

2.1 Introduction

Line spectral estimation, which aims at approximately inferring the frequency and coefficient parameters from a superposition of complex sinusoids embedded in white noise, is one of the fundamental problems in statistical signal processing. When the temporal and frequency domains are exchanged, this classical problem was reinterpreted as the problem of mathematical super-resolution recently [13–15]. This line of work promotes the use of a convex sparse regularizer to solve inverse problems involving spectrally sparse signals, distinguishing them from classical methods based on root finding and singular value decompositions (e.g., Prony's method, MUSIC, ESPIRIT, Matrix Pencil, etc.). The convex regularizer, a particular instance of the general atomic norms, has been shown to achieve optimal performance in signal completion [16], denoising [17], and outlier removal [18, 19]. For these signal processing tasks, either one can recover the spectral signal exactly (and hence extract the true frequencies precisely), or the error metric is defined using the signal instead of the frequency parameters. The most relevant question of the accuracy of noisy frequency estimation has been elusive. This work investigates the parameter estimation performance of super-resolution line spectral estimation using atomic norm minimization. More precisely, given noisy observations

$$y(t) = x^\star(t) + w(t), t = -n, \ldots, n \tag{2.1}$$

of a spectrally sparse signal

[1] This is a joint work with Gongguo Tang [2].

$$x^\star(t) = \sum_{\ell=1}^{k} c_\ell^\star \exp(i2\pi f_\ell^\star t), t = -n, \ldots, n \tag{2.2}$$

with unknown frequencies $T^\star = \{f_\ell^\star\}_{\ell=1}^k$ and complex amplitudes $\{c_\ell^\star\}_{\ell=1}^k$, we will derive conditions under which the atomic norm formulation will return the correct number of frequencies, and establish bounds on the frequency and coefficient estimation errors. An informal version of our main result is given in the following theorem, while a formal statement is presented in Theorem 2.2.1.

Theorem 2.1.1 (Informal)**.** *Suppose we observe $2n+1$ noisy consecutive samples $y(t) = x^\star(t) + w(t)$ of the signal* (2.2) *with $w(t)$ being i.i.d. complex Gaussian variables of mean zero and variance σ^2. If the unknown frequencies are well-separated, the Signal-to-Noise Ratio (SNR) is large, and the dynamic range of the coefficients is small, then with probability at least $1 - \frac{1}{n^2}$, solving an atomic norm regularized least-squares problem with a large enough regularization parameter will return exactly k estimated frequencies $\{f_\ell^{\text{glob}}\}_{\ell=1}^k$ and coefficients $\{c_\ell^{\text{glob}}\}_{\ell=1}^k$ that, when properly ordered, satisfy*

$$\max_{1\le\ell\le k} |c_\ell^\star||f_\ell^{\text{glob}} - f_\ell^\star| = O(\frac{\sqrt{\log n}}{n^{3/2}}\sigma), \tag{2.3}$$

$$\max_{1\le\ell\le k} |c_\ell^{\text{glob}} - c_\ell^\star| = O(\sqrt{\frac{\log n}{n}}\sigma). \tag{2.4}$$

We would like to first point out that this *frequency estimator* $\{f_\ell^{\text{glob}}\}$ given by the atomic norm regularized least-squares is asymptotically unbiased. The ℓ_1 norm minimization (atomic norm minimization is an extension of it) is usually considered biased because it pushes down the solution using the ℓ_1 norm. In the context of atomic norm minimization, the estimator for the coefficient vector is indeed biased for the same reason. However, the frequency estimator, which is of more interest, might still be unbiased since it is not pushed down by the atomic norm formulation. Indeed, our result shows that the frequency estimator is at least asymptotically unbiased.

Corollary 2.1.1. *Under the same setup as in Theorem 2.1.1, with probability at least $1 - \frac{1}{n^2}$, the frequency estimator obtained by the atomic norm regularized minimization is asymptotic unbiased.*

Proof. To see this, we note that for any i,

$$\begin{aligned}
\mathbb{E}[f_i^{\text{glob}}] - f_i^\star \le \mathbb{E}\{|f_i^{\text{glob}} - f_i^\star|\} &= \int_\Omega |f_i^{\text{glob}}(\omega) - f_i^\star(\omega)|\,\mathrm{d}\,\omega + \int_{\Omega^c} |f_i^{\text{glob}}(\omega) - f_i^\star(\omega)|\,\mathrm{d}\,\omega \\
&\le O(\frac{\sqrt{\log n}}{c_{\min}^\star n^{3/2}}\sigma) + \frac{2}{n^2} \\
&= o(\frac{1}{n}).
\end{aligned}$$

Here Ω is the high-probability sample space where our main result (2.3) holds, Ω^c is its complement space, and $c^\star_{\min}$ is defined as the smallest magnitude of $\{c^\star_\ell\}$. The second inequality follows from Eq. (2.3), $\int_\Omega \mathrm{d}\,\omega \leq 1$, $\int_{\Omega^c} \mathrm{d}\,\omega \leq \frac{1}{n^2}$, and the fact that any frequency is defined in $\mathbb{T} = [0, 1]$. Therefore, the frequency estimator is at least asymptotically unbiased. □

By the asymptotic unbiasedness of our *atomic frequency estimator* and considering that the Cramér-Rao bound (CRB) [20] can be viewed as the best squared error bound for any unbiased frequency estimators, we now compare our main result (2.3) (after taking the square) with the CRB, as well as the two most famous classical line spectral estimation methods, i.e., the MUSIC and Maximum Likelihood Estimation (MLE), in Table 2.1. We conclude that

Table 2.1: Comparison with the classical line spectral estimation methods.

Method	**Squared-Error Bound**
CRB [20]	$O(\frac{\sigma^2}{c^{\star 2}_{\min} n^3})$
MUSIC [20]	$O(\frac{\sigma^2}{T c^{\star 2}_{\min} n^3} + \frac{\sigma^4}{T c^{\star 4}_{\min} n^4})$
MLE [20]	$O(\frac{\sigma^2}{T c^{\star 2}_{\min} n^3} + \frac{\sigma^4}{T c^{\star 4}_{\min} n^4})$
This work (2.3)	$O(\frac{\sigma^2 \log n}{c^{\star 2}_{\min} n^3})$

the squared error bound of the *atomic frequency estimator* matches the CRB up to a logarithmic factor. We also note that the MUSIC and the MLE only have asymptotic mean squared error in the sense that the number of snapshots T has to be infinitely large [20]. We emphasize that our results are non-asymptotic, which hold for finite-length, single-snapshot signals (i.e., $T = 1$), while classical methods such as MUSIC and MLE are not efficient (i.e., approaching CRB) even with an infinite number of snapshots, as long as the signal length n is finite.

2.2 Signal Model and Atomic Norm Regularization

This work considers the spectral estimation problem: given noisy temporal samples, how well can we estimate the locations and determine the magnitudes of spectral lines? The signal of interest $x^\star(t)$ as expressed in (2.2) is composed of only a small number of spectral spikes located in a normalized interval $\mathbb{T} = [0, 1]$. We abuse notation and call $T^\star = \{f^\star_\ell\}_{\ell=1}^k$ the support of $\mathbf{x}^\star$. The number of frequencies, k, is referred to as the model order. The goal is to approximately localize these parameters from a small number $2n + 1$ of equispaced noisy samples given in (2.1). For technical simplicity, we assume $n = 2M$ is an even number. The noise components $w(t)$ are i.i.d. centrally symmetric complex Gaussian variables with variance σ^2. To simplify notation, we stack the temporal samples into vectors and write the observation model as

$$\mathbf{y} = \mathbf{x}^\star + \mathbf{w}, \tag{2.5}$$

where $\mathbf{x}^\star := [x^\star(-n), \ldots, x^\star(n)]^T$, $\mathbf{y} := [y(-n), \ldots, y(n)]^T$ and $\mathbf{w}^\star := [w^\star(-n), \ldots, w^\star(n)]^T$.

To estimate the frequency vector $\mathbf{f}^\star := [f_1^\star, \ldots, f_k^\star]^T$ and the complex coefficient vector $\mathbf{c}^\star := [c_1^\star, \ldots, c_k^\star]^T$, we assume k is small and treat $\mathbf{x}^\star$ as a sparse combination of atoms $\mathbf{a}(f) := [e^{i2\pi(-n)f}, \ldots, e^{i2\pi nf}]^T$ parameterized by frequency $f \in \mathbb{T}$, that is,

$$\mathbf{x}^\star = \sum_{\ell=1}^{k} c_\ell^\star \mathbf{a}(f_\ell^\star). \tag{2.6}$$

To exploit the structure of $\mathbf{x}^\star$ encoded in the set of atoms $\mathcal{A} := \{\mathbf{a}(f), f \in \mathbb{T}\}$, we follow [16, 21] and define the associated atomic norm as

$$\|\mathbf{x}\|_{\mathcal{A}} = \inf\left\{\sum_{\ell} |c_\ell| : \mathbf{x} = \sum_{\ell} c_\ell \mathbf{a}(f_\ell), \forall f_\ell \in \mathbb{T}, c_\ell \in \mathbb{C}\right\}. \tag{2.7}$$

The dual norm of the atomic norm, which is useful both algorithmically and theoretically, is defined for any vector $\mathbf{z}$ as $\|\mathbf{z}\|_{\mathcal{A}}^* = \sup_{f\in\mathbb{T}} |\mathbf{a}(f)^H \mathbf{z}|$, where ^{H} denotes the Hermitian (conjugate transpose) operation. To solve atomic norm minimizations numerically, the authors of [17, 22] (see also [13]) first proposed to reformulate the atomic norm (2.7) as an equivalent semidefinite program. Other numerical schemes are studied in [23–26].

Given the noisy observation model (2.5), it is natural to denoise $\mathbf{x}^\star$ by solving the atomic norm regularized minimization program [17, 22]:

$$\mathbf{x}^{\text{glob}} = \operatorname*{argmin}_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_{\mathbf{Z}}^2 + \lambda\|\mathbf{x}\|_{\mathcal{A}}. \tag{2.8}$$

For technical reasons, we used a weighted ℓ_2 norm, $\|\mathbf{z}\|_{\mathbf{Z}} := \sqrt{\mathbf{z}^H \mathbf{Z}\mathbf{z}}$, to measure data fidelity. Here $\mathbf{Z} = \text{diag}(\frac{g_M(\ell)}{M}) \in \mathbb{R}^{(4M+1)\times(4M+1)}$ with $g_M(\ell), \ell = -2M, \ldots, 2M$ defined in [16] as the discrete convolution of two triangular functions. We remark that, in practice, both a standard ℓ_2 norm $\|\cdot\|_2$ and a weighted ℓ_2 norm $\|\cdot\|_{\mathbf{Z}}$ achieve similarly satisfying performance. In this work, we use $\|\cdot\|_{\mathbf{Z}}$ with $\mathbf{Z} = \text{diag}(\frac{g_M(\ell)}{M})$ mainly for the purpose of introducing the Jackson kernel $K(f_2 - f_1) := \mathbf{a}(f_1)^H \mathbf{Z}\mathbf{a}(f_2)$ so that we can exploit the beautiful decaying properties of the Jackson kernel (see Section A.3 for more details). When we exchange the frequency and temporal domains, this weighting scheme trusts low-frequency samples more than high-frequency ones, even though the noise levels are the same. The second term is a regularization term that penalizes solutions with large atomic norms, which typically correspond to spectrally dense signals. The regularization parameter λ, whose value will be given later, controls the trade-off between data fidelity and sparsity.

Once $\mathbf{x}^{\text{glob}}$ was solved, we can extract estimates of the frequencies either from the primal optimal solution $\mathbf{x}^{\text{glob}}$ or from the corresponding dual optimal solution. Our goal is to characterize conditions such that i) we obtain exactly k estimated frequencies; ii) there is a natural correspondence between the estimated frequencies and the true frequencies, whose distances can be explicitly controlled; iii) the distances between the corresponding coefficients can also be

explicitly bounded.

To formally present the main theorem, we need to define a few more quantities. It is known that there is a resolution limit of the atomic norm approach in resolving the atoms, or the frequency parameter $\mathbf{f}^\star$, even from the noiseless data [27]. Therefore, to recover the support of the line spectral signal $\mathbf{x}^\star$, we need to impose certain separation condition on the distances of the true frequencies. For this purpose, we define $\Delta(T) = \min_{\{f_\ell, f_m\} \subset T: f_\ell \neq f_m} |f_\ell - f_m|$, where $|\cdot|$ is understood as the wrap-around distance in $\mathbb{T}$. For example, $|0.1 - 0.9| = 0.2$ under this distance. We also define 1) the dynamic range of the coefficients $B^\star := \frac{c^\star_{\max}}{c^\star_{\min}}$, where $c^\star_{\max}$ and $c^\star_{\min}$ denote the maximal and minimal modules of $\{c^\star_\ell\}_{\ell=1}^k$; 2) the normalized noise level $\gamma_0 := \sigma\sqrt{\frac{\log n}{n}}$; 3) the Noise-to-Signal Ratio $\gamma := \gamma_0 / c^\star_{\min}$ and 4) the regularization parameter $\lambda = 0.646 X^\star \gamma_0$ for some positive constant $X^\star$ to be determined later. Now we are ready to present our main result.

Theorem 2.2.1. *Suppose we observe $2n+1$ noisy consecutive samples $y_\ell = x^\star_\ell + w_\ell$ of the signal* (2.2) *or* (2.6) *with w_ℓ being i.i.d. complex Gaussian valuables of mean zero and variance σ^2. We assume $n \geq 130$ and*

$$\Delta(T^\star) \geq 2.5009/n, \tag{2.9}$$

$$X^\star B^\star \gamma \leq 10^{-3} \text{ and } B^\star / X^\star \leq 10^{-4}. \tag{2.10}$$

Then with probability at least $1 - \frac{1}{n^2}$, the optimal solution of (2.8) *has a decomposition $\mathbf{x}^{\text{glob}} = \sum_{\ell=1}^k c_\ell^{\text{glob}} \mathbf{a}(f_\ell^{\text{glob}})$ involving exactly k atoms, whose frequencies and coefficients, when properly ordered, satisfy*

$$\max_{1 \leq \ell \leq k} |c^\star_\ell| |f^{\text{glob}}_\ell - f^\star_\ell| \leq 0.4(X^\star + 35.2)\gamma_0 / n, \tag{2.11}$$

$$\max_{1 \leq \ell \leq k} |c^{\text{glob}}_\ell - c^\star_\ell| \leq (X^\star + 35.2)\gamma_0. \tag{2.12}$$

Several remarks on the conditions follow. Because of the weighting scheme we use in (2.8), our choice of λ differs from the standard one in [22] by a factor $1/n$ and ensures that the weighted dual atomic norm of the noise, $\|\mathbf{Z}\mathbf{w}\|^*_{\mathcal{A}}$, is less than λ with high probability. For technical reasons, our separation condition (2.9) is stronger compared with the previous works [13, 14, 17, 28–31][2]. The conditions (2.10) wrap several requirements on the problem parameters for the conclusions to hold: the dynamic range of the coefficients $B^\star$, the Noise-to-Signal Ratio γ, and the normalized noise γ_0 should all be small while the regularization parameter λ should be large enough as measured by $X^\star$.

It is worth noting that (2.10) implicitly imposes a strong assumption on the Noise-to-Signal Ratio

[2]Note that our separation condition is a bit larger when comparing to these recent works in super-resolution, while there are two other things to be considered. One thing is that most of these works require strong assumptions on the noise in their models (e.g., the noise is bounded), while our work removes such assumptions and hence can deal with the more general Gaussian noise. To make this possible, we have to develop a new proof strategy involving the two-step construction process of the dual certificate. Another thing is that although some prior works achieve small resolution limit (even comparable to the Relay diffraction limit [31]), they study a different problem. For example, [31] considers the signal denoising problem, that is, stable recovery of the whole signal $\mathbf{x}$ rather than the parameter estimation (i.e., the source location recovery). While the focus of our work is the accuracy of parameter estimation in Gaussian noise, which might be more significant for practical applications such as Radar and single-molecule microscopy, where precisely locating each target/point source is extremely important. Since the parameter estimation problem is much harder than the denoising problem, we have to relax a bit the separation condition for ease of analysis.

$$\gamma \leq 10^{-7}/B^{\star 2}$$

implying a sufficiently large n (but still finite). For high-level ideas, there might be two reasons to account for this phenomenon. One is that the problem of line spectral estimation is known to be sensitive to noise. Another is inherently from our proof regime, which makes the constants in Eq. (2.10) a bit conservative. More precisely, the ultimate objective is to show the boundedness and interpolation property of the target polynomial (see Proposition 2.4.1 for more details). Our method is using an "existing" dual polynomial in [13] satisfying this property and showing the distance between these two polynomials is sufficiently small. So, we require the noise level to be small, since we will see in Lemma 2.4.2 that the noise level will influence this distance.

One more remark is that the quantity $35.2\gamma_0$ in our results is related to the expected dual atomic norm of the weighted Gaussian noise $\mathbb{E}\|\mathbf{Zw}\|_{\mathcal{A}}^*$. By noting the definition $\lambda = 0.646X^\star\gamma_0$, we can rewrite the error bounds (2.11) and (2.12) in a more concise way:

$$\max_{1\leq \ell \leq k} |c_\ell^\star||f_\ell^{\text{glob}} - f_\ell^\star| = O\left(\lambda + \mathbb{E}\|\mathbf{Zw}\|_{\mathcal{A}}^*\right)/n, \tag{2.13}$$

$$\max_{1\leq \ell \leq k} |c_\ell^{\text{glob}} - c_\ell^\star| = O\left(\lambda + \mathbb{E}\|\mathbf{Zw}\|_{\mathcal{A}}^*\right). \tag{2.14}$$

Eq. (2.13) and (2.14) imply that the error bounds are determined jointly by the regularization parameter λ and the expected dual atomic norm of the weighted Gaussian noise $\mathbb{E}\|\mathbf{Zw}\|_{\mathcal{A}}^*$. Since the regularization parameter λ has the same order as $\mathbb{E}\|\mathbf{Zw}\|_{\mathcal{A}}$, the estimated frequencies and coefficients are guaranteed to have errors of orders $O\left(\mathbb{E}\|\mathbf{Zw}\|_{\mathcal{A}}^*/n\right)$ and $O\left(\mathbb{E}\|\mathbf{Zw}\|_{\mathcal{A}}^*\right)$, respectively. Remarkably, using atomic dual norm strategy allows us to deal with the Gaussian noise, while most prior works [14, 28–30] in approximate support recovery have to build their theoretical foundations on the bounded-noise assumption, which dramatically narrow down the applications.

Now we summarize the above comparisons of our result with those state-of-the-art modern support recovery methods in the Table 2.2.

Finally, our proof for Theorem 2.2.1 also reveals the connection between the atomic norm minimization (2.8) and the following ℓ_1-norm regularized, nonlinear and nonconvex least-squares program:

$$\underset{\mathbf{f},\mathbf{c}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{y}\|_{\mathbf{Z}}^2 + \lambda\|\mathbf{c}\|_1, \tag{2.15}$$

where $\mathbf{f} := [f_1, \ldots, f_k]^T$, $\mathbf{c} := [c_1, \ldots, c_k]^T$, and $\mathbf{A}(\mathbf{f}) := [\mathbf{a}(f_1), \ldots, \mathbf{a}(f_k)]$. The program (2.15) is highly nonconvex, with numerous local minima and saddle points, so solving it to global optimality is very difficult. Our analysis shows that, under the conditions of Theorem 2.2.1, the convex program (2.8) shares the same global optimum as the nonconvex program (2.15), implying that the atomic norm minimization provides a new convex way to solve the nonconvex program to global optimality. We summarize the result in the following corollary, with the formal proof listed

Table 2.2: Comparison with other modern line spectral estimation/super-resolution methods. The *Positive Measure* column refers to whether the result requires the ground-truth measure to be positive. RRC is short for Rayleigh Regularity condition [31, Definition 1.1], which generalizes the standard separation condition to clustered support. NDSC stands for the non-degenerate source condition [29, Definition 5]. In the *Support Recovery* column, *None* indicates that the work considers signal recovery instead of support recovery; *Existence* means that the work shows the existence of at least one recovered parameter around each ground-true parameter, but fails to theoretically eliminate the possibility of spurious recovered parameters; *Uniqueness* shows that around each true parameter there is one and only one recovered parameter.

Paper	**Bounded Noise**	**Positive Measure**	**Support Condition**	**SNR**	**Support Recovery**
[13, Theorem 1.5]	Yes	No	$\Delta \geq \frac{2}{n}$	Finite	None
[14, Theorem 1.2]	No	No	$\Delta \geq \frac{2}{n}$	Finite	None
[31, Theorem 1]	No	Yes	RRC	Finite	None
[28, Theorem 1.2]	Yes	No	$\Delta \geq \frac{2}{n}$	Finite	Exist
[17, Theorem 2]	No	No	$\Delta \geq \frac{2}{n}$	Finite	Exist
[29, Theorem 2]	Yes	No	NDSC	Infinite	Unique
[30, Theorem 2]	Yes	Yes	NDSC	Infinite	Unique
Theorem 2.2.1	No	No	$\Delta \geq \frac{2.5009}{n}$	Finite	Unique

in Appendix A.10.

Corollary 2.2.1. *Under the same setup as in Theorem 2.2.1, with probability at least* $1 - \frac{1}{n^2}$*, the frequencies and coefficients estimated by the atomic norm regularized minimization* (2.8) *constitute a global optimum of the* ℓ_1*-regularized nonlinear least-squares program* (2.15)*.*

2.3 Prior Art and Inspirations

Classical line spectral estimation techniques can be broadly classified into two camps: non-parametric and parametric methods. Non-parametric methods are mainly based on Fourier analysis [32, 33]. Such approaches have low computational complexities and no need for signal models. These methods have limited frequency resolution due to spectral leakage. Parametric methods, however, can achieve high resolution for parameter estimation. For example, Prony's method based on polynomial root-finding [34, 35] can resolve arbitrarily close frequencies in the noiseless setting. Yet this method is highly sensitive to noise and would fail even in the small noise regime. As stable versions of Prony's method, the subspace methods recast the noise-sensitive polynomial root-finding problem into more robust matrix eigenvalue problems. For instance, the matrix pencil method [36] arranges the observations into a matrix pencil whose generalized eigenvalues and eigenvectors contain information about the frequencies; the MUSIC algorithm [37] and the ESPRIT method [38] decompose the autocorrelation matrix into noise-subspace and signal subspace using eigenvalue decomposition and extract frequency estimates from the signal subspace. Both algorithms were shown to achieve CRB asymptotically [20, 39] when the signal length $2n + 1$ and the number of snapshots approach infinite. However, these classical methods are not efficient (i.e., approaching the CRB) even with an infinite number of snapshots, as long as the signal length is finite. Also, all classical parametric methods require knowledge

of the model order.

Modern convex optimization based methods formulate line spectral estimation as a linear inverse problem and exploit signal sparsity using ℓ_1-type regularizations. Such methods are modular, robust, and do not require knowledge of model orders. To apply the ℓ_1 regularization techniques, the continuous frequency domain is divided into a grid of discrete frequencies. When the true frequencies fall onto the discrete Fourier grid, work in compressive sensing guarantees optimal recovery performance [40–42]. When the frequencies do not fall onto the Fourier grid, however, the performance of ℓ_1 minimization degrades significantly due to basis mismatch [43]. The basis mismatch issue can be mitigated by employing finer grids [44, 45], which unfortunately often leads to numerical instability.

Atomic norm regularization avoids basis mismatch by enforcing sparsity directly in the continuous frequency domain. Given a set of atoms, possibly indexed by continuous parameters, one constructs an atomic norm in a principled way as a generalization of the ℓ_1-norm to promote signals with parsimonious representations. Using the notion of descent cones, the authors of [46] argued that the atomic norm is the best possible convex proxy for recovering sparse models. For the special line spectral estimation problem, where the atomic norm is induced by the set of parameterized complex exponentials, atomic regularizations have been shown to achieve optimal performance for several signal processing tasks. For instance, atomic norm minimization recovers a spectrally sparse signal from a minimal number of random signal samples [16], identifies and removes a maximal number of outliers [18, 19], and performs denoising with an error approaching the minimax rate [17]. When multiple measurement vectors are available, a method of exploiting the joint sparsity pattern of different signals to further improve estimation accuracy is proposed in [47–49]. All these works draw inspirations from the dual polynomial construction strategy developed in the pioneer work [13]. This work adds to this line of work by showing that the atomic framework produces optimal noisy frequency estimators.

Several closely related works also studied conditions for approximate support recovery from noisy observations. The work [28] developed error bounds on spectral support recovery for bounded noise. In [17], the authors derived suboptimal bounds for the Gaussian noise model. In [50], the authors extended this line of research to general measurement schemes beyond Fourier samples using the Beurling-LASSO (B-LASSO) program. The B-LASSO program, which minimizes a least-squares term plus the measure total variation norm, is mathematically equivalent to the atomic norm formulation. All these works [17, 28, 50] cannot guarantee the recovery of exactly one frequency in each neighborhood of the true frequencies. In this regard, the work by Duval and Peyré [29] showed that as long as the SNR is large enough and the sources are well-separated and satisfy a *non-degenerate source condition*, then total variation norm regularization can recover the correct number of the Diracs with both the coefficient error and the frequency error scale as the ℓ_2 norm of the noise. Compared with their work, our result uses the (weighted) dual atomic norm of the noise in place of the ℓ_2 norm, which differ by order of $\sqrt{n}$, allowing our bound to match the CRB up to a logarithmic factor. In addition, their work relies on a *non-degenerate source condition* [29, Definition 5] that is not proven to hold in the spectral super-resolution setting. In this sense, the present work is the first to rigorously establish

that in a high SNR regime this approach yields the right number of frequencies. Further our proof technique based on the primal-dual witness construction is also very different from that employed in [29] based on a perturbation analysis of the dual certificate in the noise-free case. In particular, our analysis reveals the connection between the convex approach and a natural nonlinear least-squares method for spectral estimation. More recently, [30] studies the support recovery for positive measures. For a comparison, there are several major differences worth remarking here: 1) in [30] more emphasis is put on the asymptotic analysis, while the presented work instead deals with non-asymptotic settings with finite signal length; 2) [30] requires the underlying noise to have finite ℓ_2 norm, which severely restricts the scope of noises satisfying such a property, excluding the well-known and most common Gaussian noise, while the presented results allow the underlying noise to be Gaussian; 3) in addition to requiring a sufficiently large signal-to-noise ratio, the main result in [30] also relies on a *non-degenerate source condition* that is not proven to hold in the spectral super-resolution setting.

2.4 Proof by Primal-Dual Witness Construction

Duality plays an important role in understanding atomic norm regularized line spectral estimation. Standard Lagrangian analysis shows that the dual problem of (2.8) has the following form:

$$\begin{aligned} \mathbf{q}^{\mathrm{glob}} = \operatorname*{argmax}_{\mathbf{q}} \ & \frac{1}{2}\|\mathbf{y}\|_{\mathbf{Z}}^2 - \frac{1}{2}\|\mathbf{y} - \lambda\mathbf{q}\|_{\mathbf{Z}}^2 \\ \text{subject to } \ & \|\mathbf{Z}\mathbf{q}\|_{\mathcal{A}}^* \leq 1. \end{aligned} \tag{2.16}$$

The complex trigonometric polynomial $Q(f) := \mathbf{a}(f)^H \mathbf{Z}\mathbf{q}$ corresponding to a dual feasible solution $\mathbf{q}$ is called a dual polynomial. The dual polynomial associated with the unique dual optimal solution $Q^{\mathrm{glob}}(f) := \mathbf{a}(f)^H \mathbf{Z}\mathbf{q}^{\mathrm{glob}}$ certifies the optimality of the unique primal optimal solution $\mathbf{x}^{\mathrm{glob}}$, and vice versa. The uniqueness of primal and dual optimal solutions is a consequence of the strong convexity of the objective functions of (2.8) and (2.16), respectively. In particular, the primal-dual optimal solutions are related by $\mathbf{q}^{\mathrm{glob}} = (\mathbf{y} - \mathbf{x}^{\mathrm{glob}})/\lambda$. We summarize these in the following proposition, with the proof given in Appendix A.9:

Proposition 2.4.1. *Let the decomposition $\hat{\mathbf{x}} = \sum_{\ell=1}^{\hat{k}} \hat{c}_\ell \mathbf{a}(\hat{f}_\ell)$ with distinct frequencies $\hat{T} = \{\hat{f}_\ell\} \subset \mathbb{T}$ and nonzero coefficients $\{\hat{c}_\ell\}$ and set $\hat{\mathbf{q}} = (\mathbf{y} - \hat{\mathbf{x}})/\lambda$. Suppose the corresponding dual polynomial $\hat{Q}(f) = \mathbf{a}(f)^H \mathbf{Z}\hat{\mathbf{q}}$ satisfies the following* Bounded Interpolation Property (BIP)*:*

$$\begin{aligned} &\hat{Q}(\hat{f}_\ell) = \operatorname{sign}(\hat{c}_\ell), \ell = 1, \ldots, \hat{k} \ \text{ (Interpolation)}; \\ &|\hat{Q}(f)| < 1, \forall f \notin \hat{T} \ \text{ (Boundedness)}; \end{aligned}$$

then $\hat{\mathbf{x}}$ and $\hat{\mathbf{q}}$ are the unique primal-dual optimal solutions to (2.8) *and* (2.16)*, that is, $\hat{\mathbf{x}} = \mathbf{x}^{\mathrm{glob}}$ and $\hat{\mathbf{q}} = \mathbf{q}^{\mathrm{glob}}$. Here the operation $\operatorname{sign}(c) := c/|c|$ for a nonzero complex number and applies entry-wise to a vector.*

Proposition 2.4.1 gives a way to extract the frequencies from the dual optimal solution – one can simply identify the frequencies where the dual polynomial corresponding to the dual optimal solution achieves magnitude 1. The uniqueness of the dual solution for (2.8) makes the construction of a dual certificate much harder compared with the line spectral signal completion problem [16] and demixing problem [18, 19]. For the latter two problems, while the primal optimal solution is unique, the dual optimal solutions are non-unique. One usually chooses one dual solution that is easier to analyze (e.g., the one with minimal energy). For the support recovery problem, we need to simultaneously construct the primal and dual solutions, which witness the optimality of each other. In the compressive sensing literature, this construction process is called the *primal-dual witness construction* [51]. In sparse recovery problems, a candidate primal solution is relatively easy to find, since when the noise is relatively small, the support of the recovered signal would not change. So one only needs to solve a LASSO problem restricted to the true support to determine the candidate coefficients, as was done in [51]. For the optimization (2.8), due to the continuous nature of the atoms, even a bit of noise would drive the support away from the true one. So to construct a candidate primal solution (hence a candidate dual solution), we need to simultaneously seek for the candidate support $\{\hat{f}_\ell\}$ and the candidate coefficients $\{\hat{c}_\ell\}$.

2.4.1 Proof Outline

We use the ℓ_1-regularized, nonlinear and nonconvex program (2.15), which we copy below, to find plausible candidates for $\{\hat{f}_\ell\}$ and $\{\hat{c}_\ell\}$:

$$\underset{\mathbf{f},\mathbf{c}}{\text{minimize}}\, \frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c}-\mathbf{y}\|_{\mathbf{Z}}^2+\lambda\|\mathbf{c}\|_1,$$

where $\mathbf{f}=[f_1,\ldots,f_k]^T$, $\mathbf{c}=[c_1,\ldots,c_k]^T$ and $\mathbf{A}(\mathbf{f})=[\mathbf{a}(f_1),\ldots,\mathbf{a}(f_k)]$. Note that we have effectively fixed the number of estimated frequencies $\hat{k}$ in Proposition 2.4.1 to be k. But unlike in compressive sensing we cannot fix $\mathbf{f}=\mathbf{f}^\star$ to solve for $\mathbf{c}$ only as was done in [51]. The program (2.15) is highly nonconvex, with numerous local minima, local maxima, and saddle points. So solving it to global optimality is hard even in theory. We are primarily interested in its local minimum $(\{\hat{f}_\ell\},\{\hat{c}_\ell\})$ in a neighborhood of the true frequencies and coefficients $(\mathbf{f}^\star,\mathbf{c}^\star)$. To find this local minimum, we will run gradient descent to (2.15) using $(\mathbf{f}^\star,\mathbf{c}^\star)$ as initialization. We will argue that under conditions presented in Theorem 2.2.1, each $\hat{f}_\ell$ and $\hat{c}_\ell$ stay close to $f_\ell^\star$ and $c_\ell^\star$ as given in (2.11) and (2.12), respectively. The major tool we use is the contraction mapping theorem. As shown in Corollary 2.2.1, the local minimum found in this manner is actually a global optimum of (2.15).

The rest of arguments consist of showing that $\hat{\mathbf{x}}=\sum_{\ell=1}^k \hat{c}_\ell\mathbf{a}(\hat{f}_\ell)$ with $\{\hat{f}_\ell\}$ and $\{\hat{c}_\ell\}$ constructed as described above satisfies the Bounded Interpolation Property of Proposition 2.4.1. The Interpolation property is automatically satisfied due to the construction process and the main challenge is to show the Boundedness property $|\hat{Q}(f)|<1, \forall f\notin \hat{T}$. The harder part is showing the Boundedness property. For ease of interpretation we first collect the definitions of

the most important variables that will be used throughout the proof, and then introduce the *main logic* and the *two-step construction process* of the proof.

Table 2.3: Notations.

Symbol	Definition
$(\mathbf{f}^\lambda, \mathbf{c}^\lambda)$	The local minima of $\mathrm{minimize}_{\mathbf{f},\mathbf{c}} \frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{x}^\star\|_{\mathbf{Z}}^2 + \lambda\|\mathbf{c}\|_1$ that is closest to $(\mathbf{f}^\star, \mathbf{c}^\star)$
$(\hat{\mathbf{f}}, \hat{\mathbf{c}})$	The local minima of $\mathrm{minimize}_{\mathbf{f},\mathbf{c}} \frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{y}\|_{\mathbf{Z}}^2 + \lambda\|\mathbf{c}\|_1$ that is closest to $(\mathbf{f}^\lambda, \mathbf{c}^\lambda)$
$\mathbf{x}^\lambda$	The primal solution defined by the local minima $(\mathbf{f}^\lambda, \mathbf{c}^\lambda)$ via $\mathbf{x}^\lambda := \sum_{\ell=1}^k c_\ell^\lambda \mathbf{a}(f_\ell^\lambda)$
$\hat{\mathbf{x}}$	The primal solution defined by the local minima $(\hat{\mathbf{f}}, \hat{\mathbf{c}})$ via $\hat{\mathbf{x}} := \sum_{\ell=1}^k \hat{c}_\ell \mathbf{a}(\hat{f}_\ell)$
$\mathbf{q}^\lambda$	The dual solution corresponding to the primal solution $\mathbf{x}^\lambda$, that is, $\mathbf{q}^\lambda := (\mathbf{x}^\star - \mathbf{x}^\lambda)/\lambda$
$\hat{\mathbf{q}}$	The dual solution corresponding to the primal solution $\hat{\mathbf{x}}$, that is, $\hat{\mathbf{q}} := (\mathbf{y} - \hat{\mathbf{x}})/\lambda$
$\mathbf{q}^\star$	$\mathbf{q}^\star := \lim_{\lambda\to 0} \mathbf{q}^\lambda$, satisfying the Boundedness and Interpolation property for $(\mathbf{f}^\star, \mathbf{c}^\star)$

Main Logic: Firstly, identifying that $Q^\star(f) := \mathbf{a}(f)^H \mathbf{Z}\mathbf{q}^\star$ satisfies the Boundedness property with some similar arguments used in [13]. Secondly, establishing that $\hat{\mathbf{q}}$ and $\mathbf{q}^\star$ are sufficiently close (so are $\hat{Q}(f) := \mathbf{a}(f)^H \mathbf{Z}\hat{\mathbf{q}}$ and $Q^\star(f) = \mathbf{a}(f)^H \mathbf{Z}\mathbf{q}^\star$). Therefore $\hat{Q}(f)$ also satisfies the Boundedness property. It turns out that directly showing the closeness of $\hat{\mathbf{q}}$ and $\mathbf{q}^\star$ is difficult. That is why we introduce the intermediate dual variable $\mathbf{q}^\lambda$ and use the *two-step construction process*, i.e., first showing $\mathbf{q}^\star$ is close to $\mathbf{q}^\lambda$ and then showing $\mathbf{q}^\lambda$ is close to $\hat{\mathbf{q}}$.

Two-step Construction Process: We will first find a local minimum $(\mathbf{f}^\lambda, \mathbf{c}^\lambda)$ of $\frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{x}^\star\|_{\mathbf{Z}}^2 + \lambda\|\mathbf{c}\|_1$ around $(\mathbf{f}^\star, \mathbf{c}^\star)$, where one should note we replaced the noisy signal $\mathbf{y}$ in (2.15) by the noise-free signal $\mathbf{x}^\star$. We will then run gradient descent to (2.15) using $(\mathbf{f}^\lambda, \mathbf{c}^\lambda)$ as initialization. The intermediate quantities $(\mathbf{f}^\lambda, \mathbf{c}^\lambda)$ will serve as a bridge between $(\mathbf{f}^\star, \mathbf{c}^\star)$ and $(\hat{\mathbf{f}}, \hat{\mathbf{c}})$ to make the proof easier. The key is noting that $\hat{Q}(f) = \mathbf{a}(f)^H \mathbf{Z}\hat{\mathbf{q}}$ is close to $Q^\lambda(f) = \mathbf{a}(f)^H \mathbf{Z}\mathbf{q}^\lambda$, where $\mathbf{q}^\lambda = (\mathbf{x}^\star - \mathbf{x}^\lambda)/\lambda$ and $\mathbf{x}^\lambda = \sum_{\ell=1}^k c_\ell^\lambda \mathbf{a}(f_\ell^\lambda)$, and $Q^\lambda(f)$ is close to $Q^\star(f) = \mathbf{a}(f)^H \mathbf{Z}\mathbf{q}^\star$. Here $\mathbf{q}^\star = \lim_{\lambda\to 0} \mathbf{q}^\lambda$ is a dual certificate used to certify the atomic decomposition of $\mathbf{x}^\star$. The former claim can be showed using the closeness of $(\mathbf{f}^\lambda, \mathbf{c}^\lambda)$ and $(\hat{\mathbf{f}}, \hat{\mathbf{c}})$. The later claim, however, must take advantage of the fact that $\mathbf{q}^\star = \lim_{\lambda\to 0} \mathbf{q}^\lambda = -\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathbf{x}^\lambda|_{\lambda=0}$ and apply the triangle inequality to

$$Q^\lambda(f) - Q^\star(f) = \frac{1}{\lambda}\int_0^\lambda \mathbf{a}(f)^H \mathbf{Z}\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^0 - \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^t\right)\mathrm{d}t,$$

where $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^0 = \lim_{\lambda\to 0}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^\lambda := \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^\star$. The closeness of $(\mathbf{f}^\lambda, \mathbf{c}^\lambda)$ and $(\mathbf{f}^\star, \mathbf{c}^\star)$ ensures that the derivatives in the integrand are also close. Finally, we exploit the properties of $Q^\star(f)$ which are similar to those established in [13] to complete the proof.

2.4.2 A Formal Proof: Applying the Contraction Mapping Theorem

Theorem 2.4.1 (Contraction Mapping Theorem)**.** *Given a Banach space $\mathcal{B}$ equipped with a norm $\|\cdot\|$, a bounded closed set $\mathcal{N} \subset \mathcal{B}$ and a map $\Theta : \mathcal{N} \to \mathcal{B}$, if $\Theta(\mathcal{N}) \subset \mathcal{N}$ (the non-escaping property) and there exists $\rho \in (0, 1)$ such*

that $\|\Theta(\mathbf{v}) - \Theta(\mathbf{w})\| \leq \rho\|\mathbf{v} - \mathbf{w}\|$ *for each* $\mathbf{v}, \mathbf{w} \in \mathcal{N}$ *(the contraction property), then there exists a unique* $\mathbf{v}^\star \in \mathcal{N}$ *such that* $\Theta(\mathbf{v}^\star) = \mathbf{v}^\star$.

This classical result helps to find a candidate solution for the construction of a valid dual certificate. To see this we first choose the bounded closed set $\mathcal{N}$ to be a small region around the target joint frequency-coefficient vector $\boldsymbol{\theta}^\star := (\mathbf{f}^\star, \mathbf{u}^\star, \mathbf{v}^\star)$ (where $\mathbf{u}^\star$ and $\mathbf{v}^\star$ denote respectively the real and imaginary parts of $\mathbf{c}^\star$). Let the fixed point map Θ be the gradient map of (2.15). The key is to determine the size of $\mathcal{N}$ in which the non-escaping and the contraction properties of the fixed point map Θ hold. Then, the contraction mapping theorem implies that iteratively performing the gradient map Θ from any initial point in $\mathcal{N}$ would produce a candidate solution that still lies in $\mathcal{N}$ (by the non-escaping property) and hence is close to $\boldsymbol{\theta}^\star$ (since $\mathcal{N}$ is small). Finally relating the fixed point equation to the BIP property shows that such a candidate solution generates a valid dual certificate.

In order to apply the contraction mapping theorem to our problem, we choose the norm in Theorem 2.4.1 to be a weighted ℓ_∞ norm $\|\cdot\|_{\tilde{\infty}}$ given by $\|(\mathbf{f}, \mathbf{u}, \mathbf{v})\|_{\tilde{\infty}} := \|(\mathbf{S}\mathbf{f}, \mathbf{u}, \mathbf{v})\|_\infty$ with $\mathbf{S} := \sqrt{|K''(0)|}\,\mathrm{diag}(|\mathbf{c}^\star|)$ and $K(\cdot)$ is the Jackson kernel (refer to Appendix A.1 for an introduction). This weighted ℓ_∞ norm is used as a metric function to define the neighborhood $\mathcal{N}$ around $\boldsymbol{\theta}^\star$. The choice of the weighting matrix $\mathbf{S}$ ensures that the larger a coefficient $c_i^\star$ is, the smaller the neighborhood in the direction of the frequency f_i. In addition, since $\sqrt{|K''(0)|}$ is of order $O(n)$, the frequency neighborhood is smaller than the coefficient neighborhood by the same order. Next, we choose the fixed point map Θ to be a weighted gradient map of (2.15)

$$\Theta(\boldsymbol{\theta}) := \boldsymbol{\theta} - \mathbf{W}^\star \nabla \left(\frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{y}\|_{\mathbf{Z}}^2 + \lambda\|\mathbf{c}\|_1 \right), \tag{2.17}$$

where the gradient ∇ is taken with respect to the parameter $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{u}, \mathbf{v})$ and the weighting matrix

$$\mathbf{W}^\star = \begin{bmatrix} \mathbf{S}^{-2} & & \\ & \mathbf{I}_k & \\ & & \mathbf{I}_k \end{bmatrix}. \tag{2.18}$$

Scaling the gradient vector by $\mathbf{W}^\star$ ensures that the Jacobian matrix of the second term in (2.17) is close to the identity matrix, which makes it easier to show the contraction property of Θ.

2.4.2.1 Two-step Construction Process

As discussed in Section 2.4.1, we divide the construction process into two steps. We first analyze the fixed point map Θ^λ obtained by replacing the noisy observation vector $\mathbf{y}$ in (2.17) by the noise-free signal $\mathbf{x}^\star$. We determine a region around $\boldsymbol{\theta}^\star$, say $\mathcal{N}^\star$, such that both the contraction and non-escaping properties of Θ^λ are satisfied in $\mathcal{N}^\star$. Then by the contraction mapping theorem, iterating the gradient map Θ^λ in $\mathcal{N}^\star$ initialized by $\boldsymbol{\theta}^\star$ generates a unique fixed point $\boldsymbol{\theta}^\lambda := (\mathbf{f}^\lambda, \mathbf{u}^\lambda, \mathbf{v}^\lambda)$. These results are summarized in the following lemma:

Lemma 2.4.1 (The First Fixed Point Map)**.** *Let the first fixed point map be the weighted gradient map of the nonconvex program* (2.15) *with the noisy signal* $\mathbf{y}$ *replaced by the noise-free signal* $\mathbf{x}^\star$*:*

$$\Theta^\lambda(\boldsymbol{\theta}) := \boldsymbol{\theta} - \mathbf{W}^\star \nabla \left(\frac{1}{2} \|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{x}^\star\|_{\mathbf{Z}}^2 + \lambda \|\mathbf{c}\|_1 \right), \tag{2.19}$$

where the gradient ∇ *is taken with respect to the parameter* $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{u}, \mathbf{v})$*. Let the regularization parameter* λ *vary in* $[0, 0.646X^\star\gamma_0]$*. Define a neighborhood* $\mathcal{N}^\star := \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_\infty \le X^\star\gamma_0/\sqrt{2}\right\}$*. Suppose that the separation condition* (2.9) *and the SNR condition* (2.10) *hold. Then the map* Θ^λ *has a unique fixed point* $\boldsymbol{\theta}^\lambda \in \mathcal{N}^\star$ *satisfying* $\Theta^\lambda(\boldsymbol{\theta}^\lambda) = \boldsymbol{\theta}^\lambda$*. Furthermore, according to the implicit function theorem,* $\boldsymbol{\theta}^\lambda$ *is a continuously differentiable function of* λ *whose derivative is given by*

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\boldsymbol{\theta}^\lambda = -(\nabla^2 G^\lambda(\boldsymbol{\theta}^\lambda))^{-1} \frac{\partial}{\partial \lambda} \nabla G^\lambda(\boldsymbol{\theta}^\lambda). \tag{2.20}$$

Finally, when λ *turns to zero, the fixed point* $\boldsymbol{\theta}^\lambda$ *converges to* $\boldsymbol{\theta}^\star$*, i.e.,* $\lim_{\lambda\to 0} \boldsymbol{\theta}^\lambda = \boldsymbol{\theta}^\star$*, and therefore* $\lim_{\lambda\to 0} \mathbf{x}^\lambda = \mathbf{x}^\star$*.*

Proof of Lemma 2.4.1. See Appendix A.4. □

We now turn to the gradient map Θ in (2.17) defined in a region $\mathcal{N}^\lambda$ around $\boldsymbol{\theta}^\lambda$. Similar to the first step, we show the contraction and non-escaping properties of Θ in $\mathcal{N}^\lambda$, which imply that iterating the gradient map Θ initialized by $\boldsymbol{\theta}^\lambda$ produces a unique fixed point $\hat{\boldsymbol{\theta}} := (\hat{\mathbf{f}}, \hat{\mathbf{u}}, \hat{\mathbf{v}})$.

Lemma 2.4.2 (The Second Fixed Point Map)**.** *Let the second fixed point map be the weighted gradient map of the nonconvex program* (2.15)*:*

$$\Theta(\boldsymbol{\theta}) = \boldsymbol{\theta} - \mathbf{W}^\star \nabla \left(\frac{1}{2} \|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{y}\|_{\mathbf{Z}}^2 + \lambda \|\mathbf{c}\|_1 \right) \tag{2.21}$$

and the region $\mathcal{N}^\lambda := \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\lambda\|_\infty \le 35.2\gamma_0/\sqrt{2}\right\}$*. Set the regularization parameter* λ *as* $0.646X^\star\gamma_0$ *in* (2.21)*. Suppose that the separation condition* (2.9) *and the SNR condition* (2.10) *hold. Then with probability at least* $1 - \frac{1}{n^2}$*,* $\Theta(\boldsymbol{\theta})$ *has a unique fixed point* $\hat{\boldsymbol{\theta}}$ *living in* $\mathcal{N}^\lambda$*.*

Proof of Lemma 2.4.2. See Appendix A.5. □

The radius of the second contraction region $\mathcal{N}^\lambda$ is determined by a high probability bound on the dual atomic norm of the Gaussian noise and ensures that $\mathcal{N}^\lambda$ is a non-escaping set for $\Theta(\boldsymbol{\theta})$. So far, we have identified the neighborhoods where the two fixed points $\boldsymbol{\theta}^\lambda$ and $\hat{\boldsymbol{\theta}}$ live in, which is the key to show the validity of the dual certificates later. Figure 2.1 illustrates the main results of Lemma 2.4.1 and Lemma 2.4.2.

Road Map. Define two pre-certificates using the two fixed points as $\mathbf{q}^\lambda := (\mathbf{x}^\star - \mathbf{x}^\lambda)/\lambda$ and $\hat{\mathbf{q}} := (\mathbf{y} - \hat{\mathbf{x}})/\lambda$ with the corresponding pre-dual polynomials denoted by $Q^\lambda(f)$ and $\hat{Q}(f)$. Here $\mathbf{x}^\lambda = \sum_{\ell=1}^k c_\ell^\lambda \mathbf{a}(f_\ell^\lambda)$ and $\hat{\mathbf{x}} =$

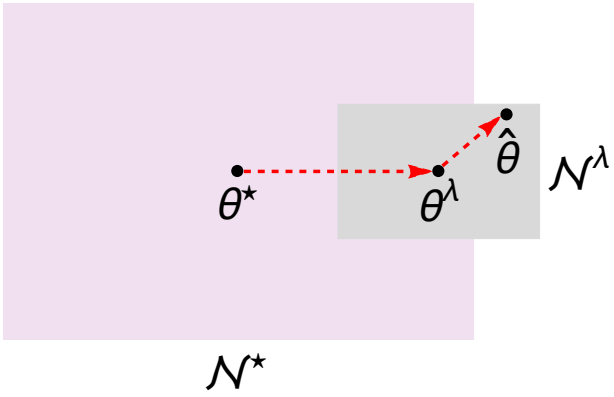


Figure 2.1: Use the true parameter vector $\boldsymbol{\theta}^\star$ as an initialization and run the first weighted gradient map (2.19) to obtain the first fixed point $\boldsymbol{\theta}^\lambda \in \mathcal{N}^\star$. Run the second weighted gradient map (2.21) initialized by $\boldsymbol{\theta}^\lambda$ to get the second fixed point $\hat{\boldsymbol{\theta}} \in \mathcal{N}^\lambda$. The closeness of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\star$ is determined by the sizes of the two neighborhoods $\mathcal{N}^\star$ and $\mathcal{N}^\lambda$, whose precise forms are given in Lemmas 2.4.1 and 2.4.2, respectively.

$\sum_{\ell=1}^k \hat{c}_\ell \mathbf{a}(\hat{f}_\ell)$. Let $\mathbf{q}^\star = \lim_{\lambda \to 0} \mathbf{q}^\lambda$. The remaining steps are to:

1. Show that $\mathbf{q}^\star$ is a valid dual certificate that certifies the atomic decomposition of $\mathbf{x}^\star$, i.e., $Q^\star(f) = \mathbf{a}(f)^H \mathbf{Z}\mathbf{q}^\star$ satisfies $Q^\star(f_\ell^\star) = \text{sign}(c_\ell^\star), \ell = 1, \ldots, k$ and $|Q^\star(f)| < 1, \forall f \notin T^\star$;

2. Use Lemma 2.4.1 to bound the pointwise distance between $Q^\star(f)$ and $Q^\lambda(f)$;

3. Use Lemma 2.4.2 to bound the pointwise distance between $Q^\lambda(f)$ and $\hat{Q}(f)$.

2.4.2.2 Showing $\mathbf{q}^\star$ is a Dual Certificate

To show that $\mathbf{q}^\star$ is a dual certificate, it is sufficient to show that $Q^\star(f)$ satisfies the Bounded Interpolation Property of Proposition 2.4.1. The Interpolation property is automatically satisfied due to the construction process, and we will show the Boundedness property using the arguments of [13]. In particular, fix an arbitrary point $f_0^\star \in T^\star$ as the reference point, and let $f_{-1}^\star$ be the first frequency in $T^\star$ that lies on the left of $f_0^\star$ while $f_1^\star$ be the first frequency in $T^\star$ that lies on the right. Here "left" and "right" are directions on the complex circle $\mathbb{T}$. We remark that the analysis depends only on the relative locations of $\{f_\ell^\star\}$. Hence, to simplify the arguments, we assume that the reference point $f_0^\star$ is at 0 by shifting the frequencies if necessary. Then we divide the region between $f_0^\star = 0$ and $f_1^\star/2$ into three parts: Near Region $\mathcal{N} := [0, 0.24/n]$, Middle Region $\mathcal{M} := [0.24/n, 0.75/n]$ and Far Region $\mathcal{F} := [0.75/n, f_1^\star/2]$. Also their symmetric counterparts are defined as $-\mathcal{N} := [-0.24/n, 0]$, $-\mathcal{M} := [-0.75/n, -0.24/n]$, and $-\mathcal{F} := [f_{-1}^\star/2, -0.75/n]$. We first show that the dual polynomial has strictly negative curvature $|Q^\star(f)|'' < 0$ in $\mathcal{N} = [0, 0.24/n]$ and $|Q^\star(f)| < 1$ in $\mathcal{M} \cup \mathcal{F} = [0.24/n, f_1^\star/2]$, implying $|Q^\star(f)| < 1$ in $\mathcal{N} \cup \mathcal{M} \cup \mathcal{F} \backslash \{f_0^\star\}$ by exploiting $|Q^\star(f_0^\star)| = 1$ and $|Q^\star(f_0^\star)|' = 0$. Then using the same symmetric arguments as in [13], we claim that $|Q^\star(f)| < 1$ in $(-\mathcal{N}) \cup (-\mathcal{M}) \cup (-\mathcal{F}) \backslash \{f_0^\star\}$. Combining these two results with the fact that the reference point $f_0^\star$ is

chosen arbitrarily from $T^\star$ (and shifted to 0), we establish that the Boundedness property of $Q^\star(f)$ holds in the entire $\mathbb{T}\backslash T^\star$.

Lemma 2.4.3 ($\mathbf{q}^\star$ is a dual certificate)**.** *The dual polynomial $Q^\star(f)$ satisfies both the Interpolation and Boundedness properties with respect to the coefficients $\{c_\ell^\star\}$ and the frequencies $\{f_\ell^\star\}$. In addition, $Q^\star(f)$ satisfies first*

$$\begin{array}{ll} Q_R^\star(f) \geq 0.887594, & Q_R^{\star\prime\prime}(f) \leq -2.24483n^2, \\ |Q_I^\star(f)| \leq 0.0183836, & |Q_I^{\star\prime\prime}(f)| \leq 0.113197n^2, \\ |Q^{\star\prime}(f)| \leq 0.821039n, & |Q^{\star\prime\prime}(f)| \leq 3.40320n^2, \end{array}$$

and

$$Q_R^\star(f)Q_R^\star(f)'' + |Q^\star(f)'|^2 + |Q_I^\star(f)||Q_I^\star(f)''| \leq -1.316313n^2 < 0$$

for $f \in \mathcal{N}$, implying $|Q^\star(f)|'' < 0$ in $\mathcal{N}$, and second,

$$\begin{array}{l} |Q^\star(f)| \leq 0.927615,\ f \in \mathcal{M}, \\ |Q^\star(f)| \leq 0.734123,\ f \in \mathcal{F}. \end{array}$$

Here the subscripts R and I denote respectively the real and imaginary parts of $Q^\star(f)$. Thus $\mathbf{q}^\star$ is a valid dual certificate to certify the atomic decomposition $\mathbf{x}^\star = \sum_{\ell=1}^k c_\ell^\star \mathbf{a}(f_\ell^\star)$ such that $\|\mathbf{x}^\star\|_{\mathcal{A}} = \sum_{\ell=1}^k |c_\ell^\star|$.

Proof of Lemma 2.4.3. See Appendix A.6. □

Next lemma, with the proof given in Appendix A.7, exploits the closeness of $\boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}^\lambda$ shown in Lemma 2.4.1 to bound the pointwise distance between $Q^\star(f)$ and $Q^\lambda(f)$.

Lemma 2.4.4 ($Q^\lambda(f)$ is close to $Q^\star(f)$)**.** *Under the settings of Lemma 2.4.1, let $Q^\lambda(f)$ and $Q^\star(f)$ be the dual polynomials corresponding to $\boldsymbol{\theta}^\lambda$ and $\boldsymbol{\theta}^\star$, respectively. Then the distances between $Q^\lambda(f)$ and $Q^\star(f)$ and their various derivatives are uniformly bounded:*

$$\begin{array}{ll} |Q^\star(f) - Q^\lambda(f)| \leq 28.7343X^\star B^\star\gamma,\ f \in \mathcal{N}, & |Q^\star(f) - Q^\lambda(f)| \leq 39.3557X^\star B^\star\gamma,\ f \in \mathcal{M}, \\ |Q^{\star\prime}(f) - Q^{\lambda\prime}(f)| \leq 44.4648nX^\star B^\star\gamma,\ f \in \mathcal{N}, & |Q^\star(f) - Q^\lambda(f)| \leq 66.1596X^\star B^\star\gamma,\ f \in \mathcal{F}, \\ |Q^{\star\prime\prime}(f) - Q^{\lambda\prime\prime}(f)| \leq 140.808n^2X^\star B^\star\gamma,\ f \in \mathcal{N}. & \end{array}$$

In the following, we will control the pointwise distance between $Q^\lambda(f)$ and $\hat{Q}(f)$ by taking advantage of the closeness of $\boldsymbol{\theta}^\lambda$ and $\hat{\boldsymbol{\theta}}$ given by Lemma 2.4.2. The key is to observe that

$$\hat{\mathbf{q}} - \mathbf{q}^\lambda = \frac{(\mathbf{y} - \hat{\mathbf{x}}) - (\mathbf{x}^\star - \mathbf{x}^\lambda)}{\lambda} = \frac{\mathbf{w}}{\lambda} + \frac{\mathbf{x}^\lambda - \hat{\mathbf{x}}}{\lambda}$$

implying

$$|Q^\lambda(f) - \hat{Q}(f)| \leq \frac{|\mathbf{a}(f)^H \mathbf{Z}\mathbf{w}|}{\lambda} + \frac{|\mathbf{a}(f)^H \mathbf{Z}(\mathbf{x}^\lambda - \hat{\mathbf{x}})|}{\lambda}. \tag{2.22}$$

This separates the distance between $Q^\lambda(f)$ and $\hat{Q}(f)$ into two parts: one is $|\mathbf{a}(f)^H \mathbf{Z}\mathbf{w}/\lambda|$ determined by the dual atomic norm of the Gaussian noise $\mathbf{w}$, which is upperbounded in Appendix A.2; the other is $|\mathbf{a}(f)^H \mathbf{Z}(\mathbf{x}^\lambda - \hat{\mathbf{x}})/\lambda|$ that can be upperbounded by the dual atomic norm of $\mathbf{x}^\lambda - \hat{\mathbf{x}}$. We summarize the final result in Lemma 2.4.5, where the proof is given in Appendix A.8.

Lemma 2.4.5 ($\hat{Q}(f)$ is close to $Q^\lambda(f)$)**.** *Under the settings of Lemma 2.4.2, let $\hat{Q}$ and Q^λ be the dual polynomials corresponding to $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\lambda$, respectively. Then the pointwise distances between $Q^\lambda(f)$ and $\hat{Q}(f)$ and their derivatives are bounded:*

$$\begin{array}{ll} |\hat{Q}(f) - Q^\lambda(f)| \leq 82.5975 B^\star/X^\star,\ f \in \mathcal{N}, & |\hat{Q}(f) - Q^\lambda(f)| \leq 114.323 B^\star/X^\star,\ f \in \mathcal{M}, \\ |\hat{Q}(f)' - Q^{\lambda'}(f)| \leq 180.283 n B^\star/X^\star,\ f \in \mathcal{N}, & |\hat{Q}(f) - Q^\lambda(f)| \leq 162.903 B^\star/X^\star,\ f \in \mathcal{F}, \\ |\hat{Q}(f)'' - Q^{\lambda''}(f)| \leq 758.404 n^2 B^\star/X^\star,\ f \in \mathcal{N}. & \end{array}$$

Proof of Theorem 2.2.1

By combining Lemmas 2.4.3, 2.4.4, and 2.4.5, we are now ready to prove Theorem 2.2.1.

Basically, we will show that $\hat{\boldsymbol{\theta}}$ constructed from the two-step gradient descent procedure and $\boldsymbol{\theta}^{\mathrm{glob}} := (\mathbf{f}^{\mathrm{glob}}, \mathbf{u}^{\mathrm{glob}}, \mathbf{v}^{\mathrm{glob}})$ are the same point. Then the error bounds follow from the closeness of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\star$. First, we show that the signal $\hat{\mathbf{x}} = \sum_{\ell=1}^k \hat{c}_\ell \mathbf{a}(\hat{f}_\ell)$ and $\hat{\mathbf{q}} = (\mathbf{y} - \hat{\mathbf{x}})/\lambda$ constructed from the second fixed point $\hat{\boldsymbol{\theta}}$ form primal and dual optimal solutions of (2.8). It suffices to show that the dual polynomial $\hat{Q}(f) = \mathbf{a}(f)^H \mathbf{Z}\hat{\mathbf{q}}$ satisfies the Bounded Interpolation Property of Proposition 2.4.1.

1) Showing the Interpolation property.

The Interpolation property has the following equivalences:

$$\begin{aligned} \hat{Q}(\hat{f}_\ell) = \mathrm{sign}(\hat{c}_\ell), \ell = 1, \ldots, k \iff & \mathbf{a}(\hat{f}_\ell)^H \mathbf{Z}(\mathbf{y} - \hat{\mathbf{x}}) = \lambda\, \mathrm{sign}(\hat{c}_\ell), \ell = 1, \ldots, k \\ \iff & \mathbf{a}(\hat{f}_\ell)^H \mathbf{Z}(\mathbf{y} - \mathbf{A}(\hat{\mathbf{f}})\hat{\mathbf{c}}) = \lambda\, \mathrm{sign}(\hat{c}_\ell), \ell = 1, \ldots, k \\ \iff & \mathbf{A}(\hat{\mathbf{f}})^H \mathbf{Z}(\mathbf{y} - \mathbf{A}(\hat{\mathbf{f}})\hat{\mathbf{c}}) = \lambda \hat{\mathbf{c}}./|\hat{\mathbf{c}}|. \end{aligned} \tag{2.23}$$

From Lemma 2.4.2, $\hat{\boldsymbol{\theta}}$ is the fixed point solution of the map $\Theta(\boldsymbol{\theta}) = \boldsymbol{\theta} - \mathbf{W}^\star \nabla G(\boldsymbol{\theta})$, i.e., $\Theta(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}$, implying $\nabla G(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ due to the invertibility of $\mathbf{W}^\star$. Invoking the explicit expression for $\nabla G(\boldsymbol{\theta})$ developed in Appendix A.3, we get

$$\nabla G(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \mathbb{R}\{(\mathbf{A}'(\hat{\mathbf{f}})\,\mathrm{diag}(\hat{\mathbf{c}}))^H \mathbf{Z}(\mathbf{A}(\hat{\mathbf{f}})\hat{\mathbf{c}} - \mathbf{y})\} \\ \mathbb{R}\{\mathbf{A}(\hat{\mathbf{f}})^H \mathbf{Z}(\mathbf{A}(\hat{\mathbf{f}})\hat{\mathbf{c}} - \mathbf{y}) + \lambda \hat{\mathbf{c}}./|\hat{\mathbf{c}}|\} \\ \mathbb{I}\{\mathbf{A}(\hat{\mathbf{f}})^H \mathbf{Z}(\mathbf{A}(\hat{\mathbf{f}})\hat{\mathbf{c}} - \mathbf{y}) + \lambda \hat{\mathbf{c}}./|\hat{\mathbf{c}}|\} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \tag{2.24}$$

Then the Interpolation property (2.23) follows from the last two row blocks of (2.24).

2) Showing the Boundedness property.

Following the same arguments preceding Lemma 2.4.3, it is sufficient to show $|\hat{Q}(f)| < 1$ in $\mathcal{N} \cup \mathcal{M} \cup \mathcal{F} \backslash \{\hat{f}_0\}$.

First, since $\hat{f}_0$ might be located in $-\mathcal{N}$ or $\mathcal{N}$, we bound $|\hat{Q}(f)|$ for $f \in (-\mathcal{N}) \cup \mathcal{N}$. The second-order Taylor expansion of $|\hat{Q}(f)|$ at $f = \hat{f}_0$ states

$$\begin{aligned} |\hat{Q}(f)| &= |\hat{Q}(\hat{f}_0)| + (f - \hat{f}_0)|\hat{Q}(\hat{f}_0)|' + \frac{1}{2}(f - \hat{f}_0)^2 |\hat{Q}(\xi)|'' \\ &= 1 + (f - \hat{f}_0)|\hat{Q}(\hat{f}_0)|' + \frac{1}{2}(f - \hat{f}_0)^2 |\hat{Q}(\xi)|'' \text{ for some } \xi \in (-\mathcal{N}) \cup \mathcal{N}, \end{aligned} \tag{2.25}$$

where for the second line we used a consequence of the interpolation property. We argue that

$$|\hat{Q}(\hat{f}_0)|' = \frac{\hat{Q}_R(\hat{f}_0)\hat{Q}_R(\hat{f}_0)' + \hat{Q}_I(\hat{f}_0)\hat{Q}_I(\hat{f}_0)'}{|\hat{Q}(\hat{f}_0)|} = \frac{\mathbb{R}\{\hat{c}_0\}\hat{Q}_R(\hat{f}_0)' + \mathbb{I}\{\hat{c}_0\}\hat{Q}_I(\hat{f}_0)'}{|\hat{c}_0||\hat{Q}(\hat{f}_0)|} = 0.$$

The last equality is a consequence of the first row block of (2.24) since $\mathbb{R}\{\hat{c}_0\}\hat{Q}_R(\hat{f}_0)' + \mathbb{I}\{\hat{c}_0\}\hat{Q}_I(\hat{f}_0)' = \mathbb{R}\{\hat{c}_0^H \mathbf{a}(\hat{f}_0)^H \mathbf{Z}(\mathbf{y} - \mathbf{A}(\hat{\mathbf{f}})\hat{\mathbf{c}})\}$. Therefore, it suffices to show that $|\hat{Q}(f)|'$ has strictly negative derivative in the symmetric Near Region $f \in (-\mathcal{N}) \cup \mathcal{N}$. By the symmetric arguments, it suffices to show this in $\mathcal{N}$. Since

$$|\hat{Q}(f)|'' = -\frac{(\hat{Q}_R(f)\hat{Q}_R(f)' + \hat{Q}_I(f)\hat{Q}_I(f)')^2}{|\hat{Q}(f)|^3} + \frac{\hat{Q}_R(f)\hat{Q}_R(f)'' + |\hat{Q}(f)'|^2 + |\hat{Q}_I(f)||\hat{Q}_I(f)''|}{|\hat{Q}(f)|},$$

we only need to show that

$$\hat{Q}_R(f)\hat{Q}_R(f)'' + |\hat{Q}(f)'|^2 + |\hat{Q}_I(f)||\hat{Q}_I(f)''| < 0,$$

which can be obtained by applying Lemma 2.4.3, Lemma 2.4.4, Lemma 2.4.5 and the triangle inequality to control these three terms $\hat{Q}_R(f)\hat{Q}_R(f)''$, $|\hat{Q}(f)'|^2$ and $|\hat{Q}_I(f)||\hat{Q}_I(f)''|$, respectively.

More precisely, the first term can be bounded by

$$\begin{aligned} &\hat{Q}_R(f)\hat{Q}_R(f)'' \\ \leq& Q_R^\star(f)Q_R^\star(f)'' + |\hat{Q}_R(f) - Q_R^\star(f)||\hat{Q}_R(f)'' - Q_R^\star(f)''| + |Q_R^\star(f)||\hat{Q}_R(f)'' - Q_R^\star(f)''| + |\hat{Q}_R(f) - Q_R^\star(f)||Q_R^\star(f)''| \\ \leq& (0.887594)(-2.24483n^2) + (28.7343X^\star B^\star\gamma + 82.5975B^\star/X^\star)(140.808n^2X^\star B^\star\gamma + 758.404n^2B^\star/X^\star) \\ &+ (1)(140.808n^2X^\star B^\star\gamma + 758.404n^2B^\star/X^\star) + (28.7343X^\star B^\star\gamma + 82.5975B^\star/X^\star)3.40320n^2 \\ \leq& -1.64194n^2, \end{aligned} \tag{2.26}$$

where we have used the SNR condition (2.10): $X^\star B^\star \gamma \le 10^{-3}, B^\star/X^\star \le 10^{-4}$ in the last line. We now bound the second term

$$\begin{aligned}|\hat{Q}(f)'|^2 =&|\hat{Q}(f)' - Q^{\star\prime}(f)|^2 + |Q^{\star\prime}(f)|^2 + 2|Q^\star(f)'||\hat{Q}(f)' - Q^{\star\prime}(f)| \\ \le&(44.4648nX^\star B^\star\gamma + 180.283nB^\star/X^\star)^2 + (0.821039n)^2 + 2(0.821039n)(44.4648nX^\star B^\star\gamma + 180.283nB^\star/X^\star) \\ \le&0.780629n^2. \end{aligned} \tag{2.27}$$

Finally, the third term can be bounded by

$$\begin{aligned}&|\hat{Q}_I(f)||\hat{Q}_I(f)''| \\ \le&(|Q_I^\star(f)| + |\hat{Q}(f) - Q^\star(f)|)(|Q_I^{\star\prime\prime}(f)| + |\hat{Q}(f)'' - Q^{\star\prime\prime}(f)|) \\ \le&(0.0183836 + (28.7343X^\star B^\star\gamma + 82.5975B^\star/X^\star))0.113197n^2 + (140.808n^2X^\star B^\star\gamma + 758.404n^2B^\star/X^\star) \\ \le&0.222917n^2. \end{aligned} \tag{2.28}$$

From (2.26), (2.27) and (2.28), we have

$$\hat{Q}_R(f)\hat{Q}_R(f)'' + |\hat{Q}(f)'|^2 + |\hat{Q}_I(f)||\hat{Q}_I(f)''| \le (-1.64194 + 0.780629 + 0.222917)n^2 < 0,$$

implying that $|\hat{Q}(f)|'' < 0$ in $\mathcal{N}$. This completes showing $|\hat{Q}(f)|'' < 0$ in $(-\mathcal{N}) \cup \mathcal{N}$ and

$$|\hat{Q}(f)| < 1, \quad \text{for } f \in (-\mathcal{N}) \cup \mathcal{N} \backslash \{\hat{f}_0\}. \tag{2.29}$$

Next, we bound $|\hat{Q}(f)|$ in Middle Region

$$\begin{aligned}|\hat{Q}(f)| \le&|Q^\star(f)| + |Q^\star(f) - Q^\lambda(f)| + |\hat{Q}(f) - Q^\lambda(f)| \\ \le&0.927615 + (39.3557X^\star B^\star\gamma + 114.323B^\star/X^\star) \\ \le&0.978403 < 1, \quad \text{for } f \in \mathcal{M}. \end{aligned} \tag{2.30}$$

Finally, we arrive at an upper bound of $|\hat{Q}(f)|$ in Far Region:

$$\begin{aligned}|\hat{Q}(f)| \le&|Q^\star(f)| + |Q^\star(f) - Q^\lambda(f)| + |\hat{Q}(f) - Q^\lambda(f)| \\ \le&0.734123 + (66.1596X^\star B^\star\gamma + 162.903B^\star/X^\star) \\ \le&0.81658 < 1, \quad \text{for } f \in \mathcal{F}. \end{aligned} \tag{2.31}$$

From (2.29), (2.30) and (2.31), we obtain that $\hat{Q}(f)$ satisfies the BIP property and hence $\hat{\mathbf{q}}$ is a valid dual certificate that certifies the optimality of $\hat{\mathbf{x}} = \sum_{\ell=1}^k \hat{c}_\ell \mathbf{a}(\hat{f}_\ell)$. The uniqueness of the decomposition as also certified by $\hat{\mathbf{q}}$ implies that $\{\hat{f}_\ell\}_{\ell=1}^k = \{f_\ell^{\text{glob}}\}_{\ell=1}^k$ and $\{\hat{c}_\ell\}_{\ell=1}^k = \{c_\ell^{\text{glob}}\}_{\ell=1}^k$, i.e., $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^{\text{glob}}$ are the same point.

As the final step, using Lemma 2.4.1, Lemma 2.4.2 and the triangle inequality, we have

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_\infty \le \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\lambda\|_\infty + \|\boldsymbol{\theta}^\lambda - \boldsymbol{\theta}^\star\|_\infty \le (X^\star + 35.2)\gamma_0/\sqrt{2}.$$

Then the desired results follow from the definition of the norm $\|\cdot\|_\infty$ and the fact that $\sqrt{|K''(0)|} \geq 3.289n^2$ for $n \geq 130$ by (A.2) and hence $1/\sqrt{2|K''(0)|} \leq 1/\sqrt{2(3.289)}/n \leq 0.3899/n \leq 0.4/n$. □

2.5 Numerical Experiments

We present numerical results to support our theoretical findings. In particular, we first examine the phase transition curve of the rate of success in Figure 2.2. In preparing Figure 2.2, k complex coefficients $c_1^\star, \ldots, c_k^\star$ were generated uniformly from the unit complex circle such that $c_{\min}^\star = c_{\max}^\star = 1$ hence $B^\star = 1$. We also generated k normalized frequencies $f_1^\star, \ldots, f_k^\star$ uniformly chosen from $[0, 1]$ such that every pair of frequencies are separated by at least $2.5/n$. Then the signal $\mathbf{x}^\star$ was formed according to (2.6). We created our observation $\mathbf{y}$ by adding Gaussian noise of mean zero and variance σ^2 to the target signal $\mathbf{x}^\star$. Let $\lambda = x\gamma_0$ (recall that $\lambda = 0.646X^\star\gamma_0$ in Theorem 2.2.1 and hence $x = 0.646X^\star$). We varied x and the Noise-to-Signal Ratio γ. For each fixed (x, γ) pair, 20 instances of the spectral line signals were generated. We then solved (2.8) for each instance and extracted the frequencies and coefficients. We declared success for an instance if i) the recovered frequency vector is within $\gamma/2n$ ℓ_∞ distance of the true frequency vector $\mathbf{f}^\star$, and ii) the recovered coefficient vector is within 2λ ℓ_∞ distance of the true frequency vector $\mathbf{c}^\star$. The rate of success for each algorithm is the proportion of successful instances.

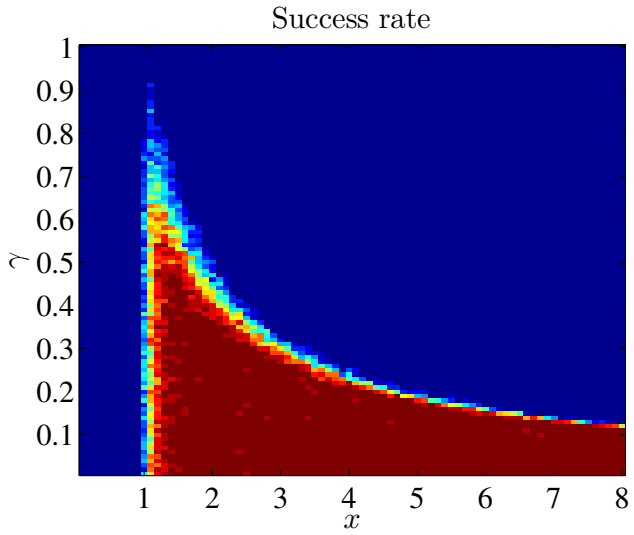


Figure 2.2: Rate of success for line spectral estimation by solving the atomic norm regularized program (2.8).

From Figure 2.2, we observe that solving (2.8) is unable to identify the sinusoidal parameters if $x \leq 1$ and the performance of the method is unstable when x is around 1. When x is set to be slightly larger than 1, however, we almost always succeed in finding good estimates of the sinusoidal parameters as long as $x\gamma \leq c$ for some small constant c. This matches the findings in Theorem 2.2.1. Figure 2.2 also shows the constants in Theorem 2.2.1 are a bit conservative.

We also run simulations to compare the mean-squared error for our frequency estimate with those for MUSIC and the MLE, as well as the CRB. The simulation results are listed in Figure 2.3. We emphasize that the MLE is initialized using the true frequencies and coefficients, which are not available in practice. We focus on the case of two unknown frequencies and examine the effect of separation. We observe that the atomic norm minimization method always outperforms MUSIC, with increased performance gap when the frequencies become closer. While the MLE performs the best, its initialization is not practical.

2.6 Conclusions

This work considers the problem of approximately estimating the frequencies and coefficients of a superposition of complex sinusoids in white noise. By using a primal-dual witness construction, we have established theoretical performance guarantees for atomic norm minimization algorithm in line spectral parameter estimation. The obtained error bounds match the Cramér-Rao lower bound up to a logarithmic factor. The relationship between resolution (separation of frequencies) and precision or accuracy of the estimator is highlighted. Our analysis also reveals that the atomic norm minimization can be viewed as a convex way to solve a ℓ_1-norm regularized, nonlinear and nonconvex least-squares problem to global optimality.

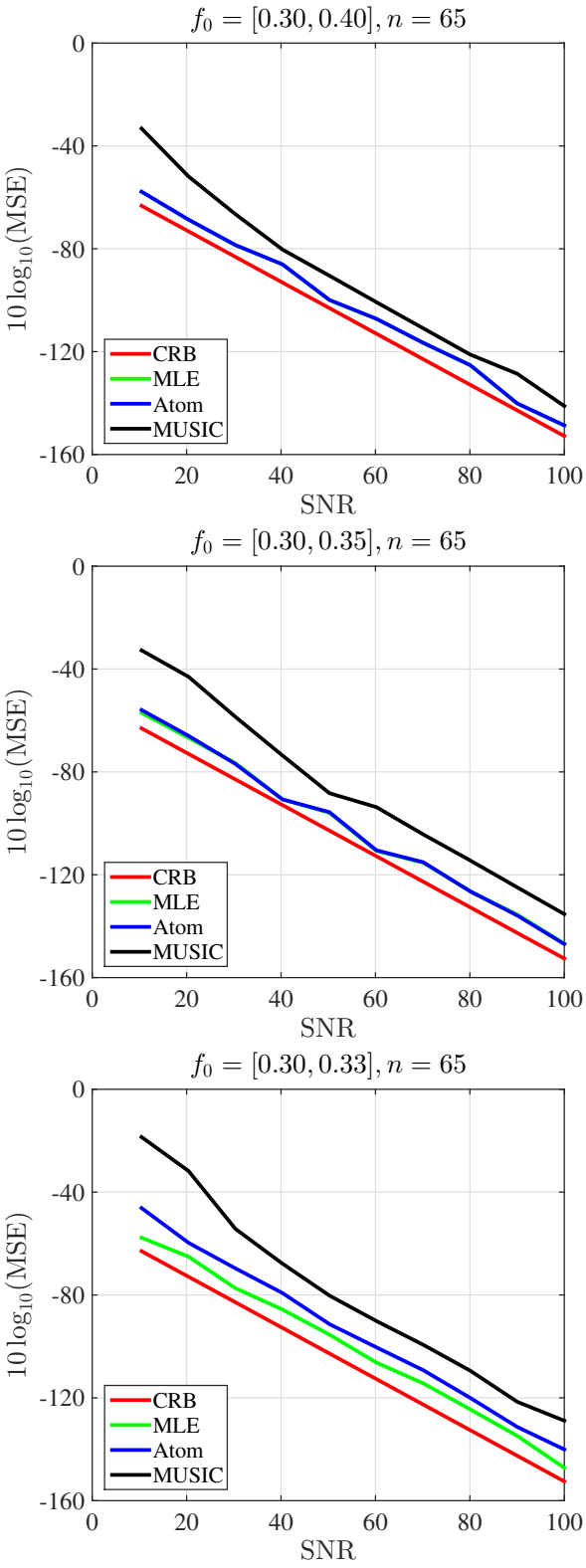


Figure 2.3: Performance comparison: Atomic norm minimization (2.8) (labeled as “Atom"), MUSIC, MLE initialized by the true parameters, and the CRB.

CHAPTER 3

A SUPER-RESOLUTION FRAMEWORK FOR TENSOR DECOMPOSITION

This work considers a super-resolution framework for overcomplete tensor decomposition. Specifically, we view tensor decomposition as a super-resolution problem of recovering a sum of Dirac measures on the sphere and solve it by minimizing a continuous analog of the ℓ_1 norm on the space of measures. The optimal value of this optimization defines the tensor nuclear norm. Similar to the separation condition in the super-resolution problem, by explicitly constructing a dual certificate, we develop incoherence conditions of the tensor factors so that they form the unique optimal solution of the continuous analog of ℓ_1 norm minimization. Remarkably, the derived incoherence conditions are satisfied with high probability by random tensor factors uniformly distributed on the sphere, implying global identifiability of random tensor factors.

3.1 Introduction

Tensors provide natural representations for massive multi-mode datasets encountered in many applications including image and video processing [52], collaborative filtering [53], array signal processing [54], convolutional networks design [55,56] and psychometrics [57]. Tensor methods also form the backbone of many machine learning, signal processing, and statistical algorithms, including independent component analysis (ICA) [58, 59], latent graphical model learning [60], dictionary learning [61], and Gaussian mixture estimation [62]. The utility of tensors in such diverse applications is mainly due to the ability to identify *overcomplete*, *non-orthogonal* factors from tensor data as already suggested by Kruskal's theorem [63]. This is known as tensor decomposition, which describes the problem of decomposing a tensor into a linear combination of a small number of rank-1 tensors. The identifiability of tensor factors is in sharp contrast to the inherent ambiguous nature of matrix decompositions without additional assumptions such as orthogonality and non-negativity.

In addition to its practical applicability, tensor decomposition is also of fundamental theoretical interest in solving linear inverse problems involving low-rank tensors. For one thing, theoretical results for tensor decomposition inform what types of rank-1 tensor combinations are identifiable given full observations. For another, a dual polynomial is constructed to certify a particular decomposition, which is useful in investigating the regularization power of the tensor nuclear norm for tensor inverse problems, including tensor completion, tensor denoising, and robust tensor principal component analysis. We expect that the *dual certificate* constructed in this work will play a role in these tensor inverse problems similar to that of the subdifferential characterization of matrix nuclear norm in matrix completion and low-rank matrix recovery [64, 65].

3.1.1 The Tensor Decomposition Inverse Problem

In this work, we focus on third-order nonsymmetric tensors that can be decomposed into a linear combination of unit-norm, rank-1 tensors of the form $\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}$, with the (i, j, k)th entry being $u_i v_j w_k$. More precisely, consider the following decomposition of a third-order nonsymmetric tensor

$$\boldsymbol{\mathcal{T}} = \sum_{p=1}^{r} \lambda_p^\star \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star. \tag{3.1}$$

Here the factors $\{(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)\}_{p=1}^r \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}^{n_3}$ might be *overcomplete* (that is, r is potentially greater than the individual tensor dimensions n_1, n_2 and n_3), *non-orthogonal* and live on the real unit spheres. Without loss of generality, we assume that the coefficients $\lambda_p^\star$ are positive as their signs can be absorbed into the factors. Tensor decomposition is the inverse problem of retrieving its rank-1 tensor factors $\{(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)\}_{p=1}^r$ from the tensor data $\boldsymbol{\mathcal{T}}$ in (3.1). After retrieving the tensor factors, finding the coefficients $\{\lambda_p^\star\}_{p=1}^r$ is simply a linear regression problem. Since the theory on complex and real tensors are very different, we emphasize that this work focuses on tensors with real entries and decompositions with real factors.

3.1.2 Our Approach

Tensor decomposition, as a generalization of the matrix singular value decomposition, is extremely challenging. First, tensor problems themselves are inherently difficult – in fact, most tensor problems are NP hard [66]. Second, we lack proper theories for basic tensor concepts and operations such as singular values, vectors, and singular value decompositions. To address these challenging issues, we view tensor decomposition as a problem of *measure estimation from moments*.

First of all, observe that retrieving the decomposition from the observed tensor entries in $\boldsymbol{\mathcal{T}}$ is equivalent to recovering a weighted sum of Dirac measures

$$\mu^\star = \sum_{p=1}^{r} \lambda_p^\star \delta(\mathbf{u} - \mathbf{u}_p^\star, \mathbf{v} - \mathbf{v}_p^\star, \mathbf{w} - \mathbf{w}_p^\star) \tag{3.2}$$

defined on the product of unit spheres $\mathbb{K} := \mathbb{S}^{n_1 - 1} \times \mathbb{S}^{n_2 - 1} \times \mathbb{S}^{n_3 - 1}$ from its third-order moments

$$\boldsymbol{\mathcal{T}} = \int_{\mathbb{K}} \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} \, \mathrm{d}\, \mu^\star.$$

In most practical scenarios, we are interested in the case where r is much smaller than the product $n_1 n_2 n_3$ (but can be significantly larger than individual dimensions n_1, n_2, and n_3). Therefore, the decomposition (3.1) is *sparse*.

Several advantages offered by this point of view are as follows. First, it provides a natural way to extend the ℓ_1 minimization in finding sparse representations for finite dictionaries [67] to tensor decomposition. By viewing the set

of rank-1 tensors $\mathcal{A} = \{\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} : (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathbb{K}\}$ as a dictionary with an infinite number of atoms, this formulation allows us to find a sparse representation of $\boldsymbol{\mathcal{T}}$ by minimizing the ℓ_1 norm of the representation coefficients with respect to the dictionary $\mathcal{A}$. More precisely, we recover $\mu^\star$ from the tensor $\boldsymbol{\mathcal{T}}$ by solving the following optimization problem

$$\underset{\mu \in \mathcal{M}(\mathbb{K})}{\text{minimize}}\, \mu(\mathbb{K}) \text{ subject to } \boldsymbol{\mathcal{T}} = \int_{\mathbb{K}} \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} \,\mathrm{d}\, \mu \tag{3.3}$$

where $\mathcal{M}(\mathbb{K})$ is the set of (nonnegative) Borel measures on $\mathbb{K}$, and $\mu(\mathbb{K})$ is the total measure/mass of the set $\mathbb{K}$ measured by the Borel measure $\mu \in \mathcal{M}(\mathbb{K})$. Second, the optimal value of the total mass minimization defines precisely the *tensor nuclear norm* [68, Proposition 3.1], which is a special case of atomic norms [21, Eq. (2)] corresponding to the atomic set $\mathcal{A}$. The tensor nuclear norm is useful in many tensor inverse problems, such as, tensor completion [52], robust tensor principal component analysis [69], and stable tensor recovery [70].

3.1.3 Main Results

The main theoretical problem investigated in this work is under what assumptions on the tensor factors $\{(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)\}_{p=1}^r$, the total mass minimization (3.3) returns the tensor decomposition (3.1). Three assumptions, namely, incoherence, bounded spectral norm, and Gram isometry, will be introduced in this work and our main result will be built upon them. For ease of exposition, in what follows, these assumptions and the main result of this work will be presented for square tensors with $n_1 = n_2 = n_3 = n$.

Assumption I: Incoherence. The tensor factors are incoherent, *i.e.,* the incoherence Δ defined below satisfies

$$\Delta := \max_{p \neq q} \max\{|\langle \mathbf{u}_p^\star, \mathbf{u}_q^\star \rangle|, |\langle \mathbf{v}_p^\star, \mathbf{v}_q^\star \rangle|, |\langle \mathbf{w}_p^\star, \mathbf{w}_q^\star \rangle|\} \leq \frac{\tau(\log n)}{\sqrt{n}}, \tag{3.4}$$

where $\tau(\cdot)$ is a polynomial function of its argument[3].

Assumption II: Bounded spectral norm. The spectral norms of $\mathbf{U} := \begin{bmatrix} \mathbf{u}_1^\star & \cdots & \mathbf{u}_r^\star \end{bmatrix}$, $\mathbf{V} := \begin{bmatrix} \mathbf{v}_1^\star & \cdots & \mathbf{v}_r^\star \end{bmatrix}$, $\mathbf{W} := \begin{bmatrix} \mathbf{w}_1^\star & \cdots & \mathbf{w}_r^\star \end{bmatrix}$ are well-controlled:

$$\max\{\|\mathbf{U}\|, \|\mathbf{V}\|, \|\mathbf{W}\|\} \leq 1 + c\sqrt{\frac{r}{n}} \tag{3.5}$$

for some constant $c > 0$.

Assumption III: Gram isometry. The Hadamard product (denoted as $\odot$) of the Gram matrices of $\mathbf{U}$ and $\mathbf{V}$ satisfies an isometric condition:

$$\|(\mathbf{U}^\top \mathbf{U}) \odot (\mathbf{V}^\top \mathbf{V}) - \mathbf{I}\| \leq \kappa(\log n) \frac{\sqrt{r}}{n}, \tag{3.6}$$

[3]Hence $\tau(\log n)$ is a polylogarithmic function of n, which is $o(n^\epsilon)$ for every exponent $\epsilon > 0$.

where $\kappa(\cdot)$ is a polynomial. Similar bounds hold for $\mathbf{U}, \mathbf{W}$, and $\mathbf{V}, \mathbf{W}$ (without loss of generality with the same polynomial $\kappa(\cdot)$).

With these assumptions, we are ready to address our theoretical problem of this work in the following theorem:

Theorem 3.1.1. *Suppose the tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{n\times n\times n}$ admits a decomposition* (3.1) *with the factors $\{(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)\}_{p=1}^r$ satisfying Assumptions I, II, III and*

$$r \leq \frac{n^{17/16}}{32c^2\sqrt{15\tau(\log n)}} \tag{3.7}$$

with the polynomial $\tau(\cdot)$ given in (3.4) *and the constant c in* (3.5)*. Then for sufficiently large n, the true factors $\{(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)\}_{p=1}^r$ can be uniquely recovered by* (3.3) *up to the sign ambiguity.*

We note that Assumptions I, II and III hold with high probability if the tensor factors $\{(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)\}_{p=1}^r$ are generated independently according to uniform distributions on the unit spheres [71, Lemmas 25, 31].

Corollary 3.1.1. *If the tensor factors $\{(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)\}_{p=1}^r$ are generated independently according to uniform distributions on the unit spheres, and if r satisfies* (3.7)*, then for sufficiently large n, solving optimization* (3.3) *is guaranteed to recover $\mu^\star$ with high probability.*

We close this section with some comments on Theorem 3.1.1 and Corollary 3.1.1.

Remark 3.1.1. Tensor decomposition using total mass minimization is an atomic decomposition problem [21, Section 2.2], which studies the conditions under which a decomposition in terms of atoms in an atomic set $\mathcal{A}$ achieves the corresponding atomic norm. For example, the singular value decomposition is an atomic decomposition for the set of unit-norm, rank-1 matrices. As shown in [27], for a large class of atomic sets, only decompositions composed of sufficiently *different* atoms are valid atomic decompositions. In particular, a necessary condition for tensor atomic decomposition is that the incoherence Δ defined in (3.4) is less than $\cos(\frac{2}{3})$ [72, Theorem 2]. However, our sufficient incoherence condition (3.4) is still significantly stronger than this necessary condition.

Remark 3.1.2. The tensor decomposition with the smallest number of rank-1 tensors is called a Canonical Polyadic (CP) decomposition and the corresponding number of rank-1 tensors is the CP-rank of the tensor, or simply the rank of the tensor. The number of factors r recovered by the optimization (3.3) may be different from the CP rank, which is called the *nuclear rank* of the tensor [68, Eq. (4.3)], and the according tensor decomposition is a *nuclear rank decomposition*.

Remark 3.1.3. Since r could be as large as $O\big(n^{17/16}/\sqrt{\tau(\log n)}\big) \gg n$ (*i.e., the number of factors r could be far more than the dimension n*), total mass minimization is guaranteed to recover *overcomplete* tensor decompositions.

Remark 3.1.4. Assumptions I-III are reasonable since they are satisfied with high probability for tensor factors uniformly lying on the unit spheres [71, Lemmas 25, 31]. Moreover, it is well-known that the incoherence for an overcom-

plete matrix $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_r] \in \mathbb{R}^{n\times r}$ with $n \leq r$ is bounded below: $\max_{p\neq q} |\langle \mathbf{u}_p, \mathbf{u}_q\rangle| \geq \sqrt{\frac{r-n}{n(r-1)}}$ [73, Chapter 1.3] and the upper bound in Assumption I is clearly larger than this lower bound for properly defined polynomial $\tau(\cdot)$.

Remark 3.1.5. The sign ambiguity is inherent in the problem formulation. In particular, we can replace the factor $(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)$ with $(a_p\mathbf{u}_p^\star, b_p\mathbf{v}_p^\star, c_p\mathbf{w}_p^\star)$ without changing the decomposition (3.1), provided that $|a_p| = |b_p| = |c_p| = 1$ and $a_p b_p c_p = 1$ (there are four such (a_p, b_p, c_p) for each p). However, this transformation gives rise to different measure representations $\mu^\star$ of the decomposition (there are 4^r of them). Therefore, the optimal solutions to (3.3) can only be unique up to this form of ambiguity.

Remark 3.1.6. It is worth commenting on the relationship between Theorem 3.1.1 and the classical Kruskal's uniqueness theorem for tensor decompositions. The Kruskal rank of matrix $\mathbf{U}$ of size $n \times r$ is defined as the maximal number $k_\mathbf{U}$ such that any $k_\mathbf{U}$ columns of $\mathbf{U}$ are linearly independent. Kruskal's theorem states that if r in the expansion (3.1) satisfies

$$r \leq \frac{1}{2}(k_\mathbf{U} + k_\mathbf{V} + k_\mathbf{W}) - 1,$$

then $\boldsymbol{\mathcal{T}}$ has a unique rank-r decomposition (up to permutation and sign ambiguities). Since the inequalities $k_\mathbf{U} \leq n$, $k_\mathbf{V} \leq n$, and $k_\mathbf{W} \leq n$ are achievable for generic matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ in $\mathbb{R}^{n\times r}$, Kruskal's theorem ensures an unique decomposition involving up to $r = \frac{3}{2}n - 1$ rank-1 (generic) factors. Note that our result holds for r up to the order $n^{17/16}$, which can be significantly larger than $\frac{3}{2}n$ for large n. Recently, the Kruskal rank r is improved to order $O(n^2)$ in [74, Corollary 6.2]. Our result on r still cannot match this bound. One might wonder whether Theorem 3.1.1 is trivial given the uniqueness of the decomposition. The caveat here is that the uniqueness holds when the decomposition involves exactly r terms, while the tensor nuclear norm, *i.e.*, the optimal value of (3.3), can potentially be achieved by decompositions involving more than r, even an infinite number of terms. In fact, the formulation takes into account decompositions with continuous supports. Theorem 3.1.1 excludes such possibility under the given conditions.

Remark 3.1.7. Corollary 3.1.1 will also be justified by numerical experiments in Section 3.5. In the experiments, we randomly sampled vectors on the unit spheres to generate the true factors of the tensor and then applied our proposed approach to decompose it. We will see that in this case, we can exactly recover the factors even for $r \gg n$.

3.1.4 Prior Art and Inspirations

Despite the advantages provided by tensor methods in many applications, their widespread adoption has been slow due to inherent computational intractability. Although the decomposition (3.1) is a multi-mode generalization of the singular value decomposition for matrices, extracting the decomposition from a given tensor is a nontrivial problem that is still under active investigation (cf. [75, 76]). Indeed, even determining the rank of a third-order tensor is an NP-hard problem [66]. A common strategy used to compute a tensor decomposition is to apply an alternating minimization

scheme. Although efficient, this approach has the drawback of not providing global convergence guarantees [75]. Recently, an approach combining alternating minimization with power iteration has gained popularity due to its ability to guarantee the tensor decomposition results under certain assumptions [71, 77].

Tensor decomposition is a special case of atomic decomposition which is to determine when a decomposition with respect to some given atomic set $\mathcal{A}$ achieves the atomic norm. For finite atomic sets, it is now well-known that if the atoms satisfy certain conditions such as the restricted isometry property, then a sparse decomposition achieves the atomic norm [78]. For the set of rank-1, unit-norm matrices, the atomic norm (the matrix nuclear norm), is achieved by orthogonal decompositions [65]. When the atoms are complex sinusoids parameterized by the frequency, Candès and Fernandez-Granda showed that atomic decomposition is solved by atoms with well-separated frequencies [13]. Similar separation conditions also show up when the atoms are translations of a known waveform [79, 80], spherical harmonics [81], and radar signals parameterized by translations and modulations [82]. Tang and Shah in [72] employed the same atomic norm idea but focused on symmetric tensors. In addition, the result of [72] does not apply to overcomplete decompositions. Under a set of conditions, including the incoherence condition ensuring the separation of tensor factors, this work characterizes a class of *nonsymmetric* and *overcomplete* tensor decompositions that achieve the tensor nuclear norm $\|\mathcal{T}\|_*$.

Another closely related line of work is matrix completion and tensor completion. Low-rank matrix completion and recovery based on the idea of nuclear norm minimization has received a great deal of attention in recent years [64, 65, 83]. A direct generalization of this approach to tensors would have been using tensor nuclear norm to perform low-rank tensor completion and recovery. However, this approach was not pursued due to the NP-hardness of computing the tensor nuclear norm [66] and the lack of analysis tools for tensor problems. The mainstream tensor completion approaches are based on various forms of matricization and application of matrix completion to the flattened tensor [52, 84, 85]. Alternating minimization can also be applied to tensor completion and recovery with performance guarantees established in recent work [86]. Most matricization and alternating minimization approaches do not yield optimal bounds on the number of measurements needed for tensor completion. One exception is [87], which used a special class of separable sampling schemes.

In contrast, we expect that the atomic norm, when specialized to tensors, will achieve the information theoretical limit for tensor completion as it does for compressive sensing, matrix completion [83], and line spectral estimation with missing data [16]. Given a set of atoms, the atomic norm is an abstraction of ℓ_1-type regularization that favors simple models. Using the notion of descent cones, Chandrasekaran et al. in [21] argued that the atomic norm is the best possible convex proxy for recovering simple models. Particularly, atomic norms are shown in many problems beyond compressive sensing and matrix completion to be able to recover simple models from minimal number of linear measurements. For example, when specialized to the atomic set formed by complex exponentials, the atomic norm can recover signals having sparse representations in the continuous frequency domain with the number of measure-

ments approaching the information theoretic limit without noise [16], as well as achieving near minimax denoising performance [17]. Continuous frequency estimation using the atomic norm is also an instance of measure estimation from (trigonometric) moments.

The rest of the chapter is organized as follows. In Section 3.2, we connect tensor decomposition to atomic decomposition, apply duality theory to derive a sufficient condition for exact decomposition, and describe extensions of the framework to tensor inverse problems. Section 3.3 presents computational methods to solve the tensor decomposition. We then proceed to develop a proof of Theorem 3.1.1 in Section 3.4. In Section 3.5, we validate our theory using numerical experiments. Additional proofs are given in the appendix.

3.2 Tensor Decomposition, Atomic Norms, and Duality

3.2.1 Tensor Decomposition as an Atomic Decomposition

In this work, we view tensor decomposition in the frameworks of both atomic norms and measure estimation. The unit sphere of $\mathbb{R}^n$ is denoted by $\mathbb{S}^{n-1}$, and the direct product of three unit spheres $\mathbb{S}^{n-1} \times \mathbb{S}^{n-1} \times \mathbb{S}^{n-1}$ by $\mathbb{K}$. The tensor atomic set is denoted by $\mathcal{A} = \{\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} : (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathbb{K}\}$ parameterized by the set $\mathbb{K}$, where $\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}$ is a rank-1 tensor with the (i, j, k)th entry being $u_i v_j w_k$. For any tensor $\boldsymbol{\mathcal{T}}$, its atomic norm with respect to $\mathcal{A}$ is defined by [21, Eq. (2)]

$$\begin{aligned} \|\boldsymbol{\mathcal{T}}\|_{\mathcal{A}} &= \inf\{t : \boldsymbol{\mathcal{T}} \in t \operatorname{conv}(\mathcal{A})\} \\ &= \inf\left\{ \sum_p \lambda_p : \boldsymbol{\mathcal{T}} = \sum_p \lambda_p \mathbf{u}_p \otimes \mathbf{v}_p \otimes \mathbf{w}_p, \lambda_p > 0, (\mathbf{u}_p, \mathbf{v}_p, \mathbf{w}_p) \in \mathbb{K} \right\}, \end{aligned} \tag{3.8}$$

where $\operatorname{conv}(\mathcal{A})$ is the convex hull of the atomic set $\mathcal{A}$, and a scalar multiplying a set scales every element in the set. Therefore, the tensor atomic norm is the minimal ℓ_1 norm of its expansion coefficients among all valid expansions in terms of unit-norm, rank-1 tensors. The atomic norm $\|\boldsymbol{\mathcal{T}}\|_{\mathcal{A}}$ defined in (3.8) is also called the tensor nuclear norm and denoted by $\|\boldsymbol{\mathcal{T}}\|_*$ in [68, Eq. (2.7)]. We will use these two names and notations interchangeably in the following. The way of defining the tensor nuclear norm is precisely the same as that of defining the matrix nuclear norm.

We argue that the two lines in the definition (3.8) are consistent and are also equivalent to (3.3) as follows. Since $\operatorname{conv}(\mathcal{A}) = \{\boldsymbol{\mathcal{T}} : \boldsymbol{\mathcal{T}} = \int_{\mathbb{K}} \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} \, \mathrm{d}\, \mu, \mu \in \mathcal{M}(\mathbb{K}), \mu(\mathbb{K}) \leq 1\}$, the first line in the definition (3.8) implies that $\|\boldsymbol{\mathcal{T}}\|_{\mathcal{A}}$ is equal to the optimal value of (3.3). Compared with the measure optimization (3.3), the feasible region of the minimization defining the atomic norm in the second line of (3.8) is restricted to discrete measures. However, these two optimizations share the same optimal value as a consequence of Carathéodory's convex hull theorem, which states that if a point $\mathbf{x} \in \mathbb{R}^d$ lies in the convex hull of a set, then $\mathbf{x}$ can be written as a convex combination of at most $d + 1$ points of that set [88, Theorem 2.3]. Since $\boldsymbol{\mathcal{T}} \in \|\boldsymbol{\mathcal{T}}\|_{\mathcal{A}} \operatorname{conv}(\mathcal{A}) = \operatorname{conv}(\|\boldsymbol{\mathcal{T}}\|_{\mathcal{A}} \mathcal{A})$, $\boldsymbol{\mathcal{T}}$ can be expressed as a convex combination of at most $n^3 + 1$ points of the set $\|\boldsymbol{\mathcal{T}}\|_{\mathcal{A}} \mathcal{A}$, implying that the optimal value is achieved by a discrete

measure with support size at most $n^3 + 1$. This argument establishes that the two lines in (3.8) as well as the measure optimization (3.3) are equivalent. Therefore, the atomic norm framework and the measure optimization framework are two different formulations of the same problem, with the former setting the stage in the finite dimensional space and the latter in the infinite-dimensional space of measures.

Given an abstract atomic set, the problem of atomic decomposition seeks the conditions under which a decomposition in terms of the given atoms achieves the atomic norm. In this sense, the tensor decomposition considered in this work is an atomic decomposition problem.

3.2.2 Duality

Duality plays an important role in analyzing atomic tensor decomposition. We again approach duality from both perspectives of atomic norms and measure estimation.

First, we find the dual problem of the optimization problem (3.3). Given $\boldsymbol{\mathcal{Q}}, \boldsymbol{\mathcal{T}} \in \mathbb{R}^{n\times n\times n}$, we define the tensor inner product $\langle \boldsymbol{\mathcal{Q}}, \boldsymbol{\mathcal{T}} \rangle := \sum_{i,j,k} Q_{ijk} T_{ijk}$. Standard Lagrangian analysis shows that the dual problem of (3.3) is the following semi-infinite program, which has an infinite number of constraints:

$$\begin{aligned} &\underset{\boldsymbol{\mathcal{Q}}\in\mathbb{R}^{n\times n\times n}}{\text{maximize}} \ \langle \boldsymbol{\mathcal{Q}}, \boldsymbol{\mathcal{T}} \rangle \\ &\text{subject to} \ \langle \boldsymbol{\mathcal{Q}}, \mathbf{u}\otimes\mathbf{v}\otimes\mathbf{w}\rangle \le 1, \forall (\mathbf{u},\mathbf{v},\mathbf{w}) \in \mathbb{K} \end{aligned} \tag{3.9}$$

The polynomial $q(\mathbf{u},\mathbf{v},\mathbf{w}) := \langle \boldsymbol{\mathcal{Q}}, \mathbf{u}\otimes\mathbf{v}\otimes\mathbf{w}\rangle = \sum_{i,j,k} Q_{ijk} u_i v_j w_k$ corresponding to a dual feasible solution $\boldsymbol{\mathcal{Q}}$ of (3.9) is called a dual polynomial. The dual polynomial associated with an optimal dual solution can be used to certify the optimality of a particular decomposition, as demonstrated by the following proposition.

Proposition 3.2.1. *Suppose the set of rank-1 tensors* $\{\mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star\}_{p=1}^r$ *given in* (3.1) *is linearly independent. If there exists a dual solution* $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{n\times n\times n}$ *to* (3.9) *such that the corresponding dual polynomial* $q : \mathbb{K} \to \mathbb{R}$

$$q(\mathbf{u},\mathbf{v},\mathbf{w}) := \langle \boldsymbol{\mathcal{Q}}, \mathbf{u}\otimes\mathbf{v}\otimes\mathbf{w}\rangle \tag{3.10}$$

satisfies the following Boundedness and Interpolation Property (BIP):

$$q(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star) = 1 \ \text{ for } p \in [r] \ \text{ (Interpolation)} \tag{3.11a}$$

$$q(\mathbf{u},\mathbf{v},\mathbf{w}) < 1 \ \text{ in } \mathbb{K} \setminus S^\star \ \text{ (Boundedness)} \tag{3.11b}$$

where $[r] := \{1,\dots,r\}$ *and*

$$S^\star := \{(a_p\mathbf{u}_p^\star, b_p\mathbf{v}_p^\star, c_p\mathbf{w}_p^\star) : |a_p| = |b_p| = |c_p| = a_p b_p c_p = 1, p \in [r]\}, \tag{3.12}$$

then $\mu^\star$ *given in* (3.2) *is the unique optimal solution to* (3.3) *up to sign ambiguity.*

Proof. In view of (3.9), any $\mathcal{Q}$ that satisfies the BIP in (3.11) is a dual feasible solution. We also have

$$\langle \mathcal{Q}, \mathcal{T} \rangle = \left\langle \mathcal{Q}, \sum_{p=1}^{r} \lambda_p^\star \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star \right\rangle = \sum_{p=1}^{r} \lambda_p^\star \langle \mathcal{Q}, \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star \rangle = \sum_{p=1}^{r} \lambda_p^\star q(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star) = \mu^\star(\mathbb{K})$$

establishing a zero-duality gap of the primal-dual feasible solution $(\mu^\star, \mathcal{Q})$. As a consequence, $\mu^\star$ is a primal optimal solution to (3.3) and $\mathcal{Q}$ is a dual optimal solution to (3.9).

For uniqueness, suppose $\hat{\mu}$ is another primal optimal solution to (3.3). If $\hat{\mu}(\mathbb{K} \setminus S^\star) > 0$, then

$$\begin{aligned} \mu^\star(\mathbb{K}) = \langle \mathcal{Q}, \mathcal{T} \rangle = \left\langle \mathcal{Q}, \int_{\mathbb{K}} \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} \, \mathrm{d}\, \hat{\mu} \right\rangle &= \sum_{(\mathbf{u},\mathbf{v},\mathbf{w}) \in S^\star} \hat{\mu}(\mathbf{u}, \mathbf{v}, \mathbf{w}) q(\mathbf{u}, \mathbf{v}, \mathbf{w}) + \int_{\mathbb{K} \setminus S^\star} q(\mathbf{u}, \mathbf{v}, \mathbf{w}) \, \mathrm{d}\, \hat{\mu} \\ &< \hat{\mu}(S^\star) + \int_{\mathbb{K} \setminus S^\star} 1 \, \mathrm{d}\, \hat{\mu} \\ &= \hat{\mu}(\mathbb{K}) \end{aligned}$$

contradicting the optimality of $\hat{\mu}$. So all optimal solutions are supported on $S^\star$. To remove the sign ambiguity, we can assume an optimal solution is supported on $\{\mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star\}_{p=1}^r$. Since $\{\mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star\}_{p=1}^r$ is linearly independent by assumption, the coefficients $\lambda_p^\star$ can be uniquely determined from solving the linear system of equations encoded in $T = \sum_{p=1}^r \lambda_p^\star \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star$. This proves the uniqueness (up to sign ambiguity). □

3.2.3 Dual Certificate and Subdifferential

The dual optimal solution $\mathcal{Q}$ satisfying the BIP is called a *dual certificate*, which is used frequently as the starting point to derive several atomic decomposition and super-resolution results [13, 16, 72, 81]. In Section 3.4, we will explicitly construct a *dual certificate* to prove Theorem 3.1.1. In this subsection, we will relate the *dual certificate* with the subdifferential of the tensor nuclear norm.

First, the dual norm of the tensor nuclear norm, *i.e.*, the tensor spectral norm, of a tensor $\mathcal{Q}$ is given by

$$\|\mathcal{Q}\| := \sup_{\mathcal{T}: \|\mathcal{T}\|_* \leq 1} \langle \mathcal{Q}, \mathcal{T} \rangle = \sup_{(\mathbf{u},\mathbf{v},\mathbf{w}) \in \mathbb{K}} \langle \mathcal{Q}, \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} \rangle.$$

The equality is due to the fact that the atomic set $\mathcal{A}$ are the extreme points of the unit nuclear norm ball $\{\mathcal{T} : \|\mathcal{T}\|_* \leq 1\}$. In light of the spectral norm definition, we rewrite the dual problem (3.9) as

$$\underset{\mathcal{Q} \in \mathbb{R}^{n \times n \times n}}{\text{maximize}} \ \langle \mathcal{Q}, \mathcal{T} \rangle \ \text{subject to} \ \|\mathcal{Q}\| \leq 1 \tag{3.13}$$

which is precisely the definition of the dual norm of the tensor spectral norm, *i.e.*, the tensor nuclear norm.

The subdifferential (the set of subgradients) of the tensor nuclear norm is defined by [73, Definition B.20]

$$\partial \|\cdot\|_*(\mathcal{T}) = \{\mathcal{Q} \in \mathbb{R}^{n \times n \times n} : \|\mathcal{R}\|_* \geq \|\mathcal{T}\|_* + \langle \mathcal{R} - \mathcal{T}, \mathcal{Q} \rangle, \text{for all } \mathcal{R} \in \mathbb{R}^{n \times n \times n}\}, \tag{3.14}$$

which has an equivalent representation [89, Section 1]

$$\partial\|\cdot\|_*(\boldsymbol{\mathcal{T}}) = \left\{\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{n\times n\times n} : \|\boldsymbol{\mathcal{T}}\|_* = \langle \boldsymbol{\mathcal{Q}}, \boldsymbol{\mathcal{T}}\rangle, \|\boldsymbol{\mathcal{Q}}\| \le 1\right\}. \tag{3.15}$$

For $\boldsymbol{\mathcal{T}}$ having an atomic decomposition given in (3.1), it can be established that the defining properties of subdifferential (3.15) are equivalent to

$$\langle \boldsymbol{\mathcal{Q}}, \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star\rangle = 1, \text{ for } p \in [r] \tag{3.16a}$$
$$\langle \boldsymbol{\mathcal{Q}}, \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}\rangle \le 1, \text{ for} (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathbb{K} \tag{3.16b}$$

We recognize that the BIP in (3.11) is a strengthened version of the subdifferential conditions (3.16). Therefore, a *dual certificate*, *i.e.,* any $\boldsymbol{\mathcal{Q}}$ satisfying the BIP, is an element of the subdifferential $\partial\|\cdot\|_*(\boldsymbol{\mathcal{T}})$. The BIP in fact means that $\boldsymbol{\mathcal{Q}}$ is an interior point of $\partial\|\cdot\|_*(\boldsymbol{\mathcal{T}})$. Our proof strategy for Theorem 3.1.1 is to construct such an interior point in Section 3.4. This is in contrast to the matrix case, for which we have an explicit characterization of the entire subdifferential of the nuclear norm using the singular value decomposition (more explicit than the one given in (3.15)). More specifically, suppose $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ is the (compact) singular value decomposition of $\mathbf{X} \in \mathbb{R}^{m\times n}$ with $\mathbf{U} \in \mathbb{R}^{m\times r}, \mathbf{V} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{\Sigma}$ being an $r \times r$ diagonal matrix. Then the subdifferential of the matrix nuclear norm at $\mathbf{X}$ is given by [65, Eq. (2.9)]

$$\partial\|\cdot\|_*(\mathbf{X}) = \{\mathbf{U}\mathbf{V}^\top + \mathbf{W} : \mathbf{U}^\top\mathbf{W} = \mathbf{0}, \mathbf{W}\mathbf{V} = \mathbf{0}, \|\mathbf{W}\| \le 1\}.$$

It is challenging to obtain such a characterization for tensors unless the tensor admits an orthogonal rank-1 decomposition.

3.2.4 Extension: Regularization Using Tensor Nuclear Norm

Independent from practical considerations, we investigate tensor decomposition for theoretical reasons. Similar to regularizing matrix inverse problems using the matrix nuclear norm, the tensor nuclear norm can be used to regularize tensor inverse problems. Suppose we observe an unknown low-rank tensor $\boldsymbol{\mathcal{T}}^\star$ through the linear measurement model $\mathbf{y} = \mathcal{B}(\boldsymbol{\mathcal{T}}^\star)$, we would like to recover the tensor $\boldsymbol{\mathcal{T}}^\star$ from the observation $\mathbf{y}$. For instance, when $\mathcal{B}$ samples the individual entries of $\boldsymbol{\mathcal{T}}^\star$, we are looking at a tensor completion problem. We propose recovering $\boldsymbol{\mathcal{T}}^\star$ by solving

$$\underset{\boldsymbol{\mathcal{T}}\in\mathbb{R}^{n\times n\times n}}{\text{minimize}} \|\boldsymbol{\mathcal{T}}\|_* \text{ subject to } \mathbf{y} = \mathcal{B}(\boldsymbol{\mathcal{T}}) \tag{3.17}$$

which favors a low-rank solution. To establish recoverability, we can construct a *dual certificate* $\boldsymbol{\mathcal{Q}}$ of the form $\mathcal{B}^*(\boldsymbol{\lambda})$, whose corresponding dual polynomial satisfies the BIP. Here $\mathcal{B}^*$ is the adjoint operator of $\mathcal{B}$. When the operator $\mathcal{B}$ is random, the concentration of measure guarantees that we can construct a *dual certificate* $\mathcal{B}^*(\boldsymbol{\lambda})$ that is close to

the one constructed in the full data case. This fact can then be exploited to verify the BIP of $\mathcal{B}^*(\boldsymbol{\lambda})$ and to establish exact recovery. When the atoms are complex exponentials parameterized by continuous frequencies, this strategy is adopted to establish the compressed sensing off the grid result (the completion problem) [16] building upon the dual polynomial constructed for the super-resolution problem (the full data case) [13]. It shows that the number of random linear measurements required for exact recovery approaches the information theoretical limit. In addition to exact recovery from noise-free measurements, the *dual certificate* for the full data case can also be utilized to derive near-minimax denoising performance [17, 22], approximate support recovery [1, 2, 28], and robust recovery from observations corrupted by outliers [19, 90]. We expect that the dual polynomial constructed for tensor decomposition will play a similar role for tensor inverse problems, enabling the development of tensor results parallel to their matrix counterparts such as matrix completion, denoising, and robust principal component analysis. We leave these as our future work.

3.3 Computational Methods

Our main theorem shows that when the tensor factors $\{(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)\}_{p=1}^r$ satisfy Assumptions I, II, III, we can recover the tensor decomposition of r up to the order of $n^{17/16}$ by solving the convex, infinite-dimensional optimization (3.3). However, as a measure optimization problem, optimization problem (3.3) is not directly solvable on a computer. In this section, we propose two computational methods, which are respectively based on:

1. The Burer-Monteiro factorization approach [6, 8, 91–93];
2. The Lasserre hierarchy [94, 95].

3.3.1 The Burer-Monteiro Factorization Approach

When dealing with convex programs involved with a large matrix variable $\mathbf{X}$, Burer and Monteiro in [91] proposed factoring the variable $\mathbf{X}$ into the product of two smaller rectangular matrices $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ and then treating them as the new optimization variables. As a typical example, Recht et al. in [65] used this approach to get that the matrix nuclear norm for any $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ equals the optimum value of the following optimization

$$\underset{\mathbf{U}\in\mathbb{R}^{n_1\times \tilde{r}}, \mathbf{V}\in\mathbb{R}^{n_2\times\tilde{r}}}{\text{minimize}} \frac{1}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right) \text{ subject to } \mathbf{X} = \mathbf{U}\mathbf{V}^\top$$

with $\tilde{r} \geq \operatorname{rank}(\mathbf{X})$. Similarly, when applying this idea to the tensor nuclear norm, we have the following result.

Proposition 3.3.1. *Suppose the decomposition that achieves the tensor nuclear norm $\|\boldsymbol{\mathcal{T}}\|_*$ involves r terms and $\tilde{r} \geq r$, then $\|\boldsymbol{\mathcal{T}}\|_*$ is equal to the optimal value of the following optimization:*

$$\underset{\{\mathbf{u}_p,\mathbf{v}_p,\mathbf{w}_p\}_{p=1}^{\tilde{r}}}{\text{minimize}} \sum_{p=1}^{\tilde{r}} \frac{1}{3}\left(\|\mathbf{u}_p\|_2^3 + \|\mathbf{v}_p\|_2^3 + \|\mathbf{w}_p\|_2^3\right)$$
$$\text{subject to } \boldsymbol{\mathcal{T}} = \sum_{p=1}^{\tilde{r}} \mathbf{u}_p \otimes \mathbf{v}_p \otimes \mathbf{w}_p \tag{3.18}$$

Proof. Suppose the tensor nuclear norm is achieved by the decomposition $\boldsymbol{\mathcal{T}} = \sum_{p=1}^{r} \lambda_p^\star \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star$. Then we note that $\{\lambda_p^{\star 1/3}\mathbf{u}_p^\star, \lambda_p^{\star 1/3}\mathbf{v}_p^\star, \lambda_p^{\star 1/3}\mathbf{w}_p^\star\}_{p=1}^{\tilde{r}}$ forms a feasible solution to (3.18) when $\tilde{r} = r$. When $\tilde{r} > r$, we can zero-pad the remaining factors $\{\mathbf{u}_p, \mathbf{v}_p, \mathbf{w}_p\}_{p=r+1}^{\tilde{r}}$. The objective function value at this feasible solution is $\frac{1}{3}(\sum_{p=1}^{\tilde{r}} 3\lambda_p^\star) = \|\boldsymbol{\mathcal{T}}\|_*$. This shows that $\|\boldsymbol{\mathcal{T}}\|_*$ is greater than the optimal value of (3.18).

To show the other direction, suppose an optimal solution of (3.18) is $\{\mathbf{u}_p, \mathbf{v}_p, \mathbf{w}_p\}_{p=1}^{\tilde{r}}$. Define $\lambda_p := \|\mathbf{u}_p\|_2\|\mathbf{v}_p\|_2\|\mathbf{w}_p\|_2$, for $p \in [\tilde{r}]$. Then,

$$\boldsymbol{\mathcal{T}} = \sum_{p:\lambda_p \neq 0} \lambda_p \frac{\mathbf{u}_p}{\|\mathbf{u}_p\|_2} \otimes \frac{\mathbf{v}_p}{\|\mathbf{v}_p\|_2} \otimes \frac{\mathbf{w}_p}{\|\mathbf{w}_p\|_2}.$$

Finally, by definition of the tensor nuclear/atomic norm (3.8), we have

$$\|\boldsymbol{\mathcal{T}}\|_* \leq \sum_{p:\lambda_p \neq 0} \lambda_p = \sum_{p=1}^{\tilde{r}} \lambda_p = \sum_{p=1}^{\tilde{r}} \|\mathbf{u}_p\|_2\|\mathbf{v}_p\|_2\|\mathbf{w}_p\|_2 \leq \frac{1}{3}\sum_{p=1}^{\tilde{r}} \left[\|\mathbf{u}_p\|_2^3 + \|\mathbf{v}_p\|_2^3 + \|\mathbf{w}_p\|_2^3\right],$$

which is the optimal value of (3.18). Therefore, the optimal value of (3.18) is equal to $\|\boldsymbol{\mathcal{T}}\|_*$. □

Proposition 3.3.1 implies that when an upper bound on r is known, we can solve the nonlinear (and non-convex) program (3.18) to compute the tensor nuclear norm (and obtain the corresponding decomposition). Numerical simulations suggest that the nonlinear program (3.18), when solved using the ADMM approach [96], has superior performance. Although in theory only local optima can be obtained for the nonlinear programming formulation (3.18), in practice for tensors with random factors, our experiments show that the decomposition can almost always be recovered by the ADMM implementation of (3.18).

Remark 3.3.1. Proposition 3.3.1 can be generalized to compute the nuclear norm of an arbitrary-sized tensor, including the matrix – a 2nd-order tensor. Basically, for a general dth-order tensor $\boldsymbol{\mathcal{T}} = \sum_{p=1}^{\tilde{r}} \mathbf{u}_p^{(1)} \otimes \cdots \otimes \mathbf{u}_p^{(d)}$, the nuclear norm $\|\boldsymbol{\mathcal{T}}\|_*$ is given by the optimum value of the following program

$$\underset{\{\mathbf{u}_p^{(1)},\cdots,\mathbf{u}_p^{(d)}\}_{p=1}^{\tilde{r}}}{\text{minimize}} \sum_{p=1}^{\tilde{r}} \frac{1}{d}\left(\|\mathbf{u}_p^{(1)}\|_2^d + \|\mathbf{u}_p^{(2)}\|_2^d + \cdots + \|\mathbf{u}_p^{(d)}\|_2^d\right)$$
$$\text{subject to } \boldsymbol{\mathcal{T}} = \sum_{p=1}^{\tilde{r}} \mathbf{u}_p^{(1)} \otimes \mathbf{u}_p^{(2)} \otimes \cdots \otimes \mathbf{u}_p^{(d)} \tag{3.19}$$

3.3.2 The Lasserre Hierarchy

As a special moment problem, the optimization problem (3.3) can be approximated increasingly tightly by the semidefinite programs in the Lasserre relaxation hierarchy [94, 95]. The Lasserre hierarchy proposes that instead of optimizing problem (3.3) with respect to the measure μ, we can equivalently optimize the (infinite-dimensional) moment sequence corresponding to μ:

$$\mathbf{m} = [m_{\boldsymbol{\alpha}}] = \int_{\mathbb{K}} \boldsymbol{\xi}^{\boldsymbol{\alpha}} \mu(\mathrm{d}\,\boldsymbol{\xi}).$$

Here the combined variable $\boldsymbol{\xi} = [\mathbf{u}^\top\ \mathbf{v}^\top\ \mathbf{w}^\top]^\top \in \mathbb{R}^{3n}$, the multi-integer index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{3n})$, and the monomial $\boldsymbol{\xi}^{\boldsymbol{\alpha}} = \xi_1^{\alpha_1}\xi_2^{\alpha_2}\cdots\xi_{3n}^{\alpha_{3n}}$. To get a finite-dimensional relaxation with a relaxation order of d, we truncate the infinite-dimensional moment sequence $\mathbf{m}$ to a finite-dimensional vector $\mathbf{m}_{2d}$ that includes moments up to order $2d$, *i.e.*, to retain moments $m_{\boldsymbol{\alpha}}$ with $|\boldsymbol{\alpha}| = \sum_{i=1}^{3n} \alpha_i \leq 2d$. Three sets of linear matrix inequalities should be satisfied for a vector $\mathbf{m}_{2d}$ to be the $2d$th-order truncation of a moment sequence on $\mathbb{K}$:

- First, since the moment matrix here is related with some positive measure μ, *i.e.*,

$$\mathbf{M}_{2d}(\mathbf{m}_{2d}) := \int_{\mathbb{K}} \begin{bmatrix} 1 \\ \xi_1 \\ \vdots \\ \xi_{3n}^d \end{bmatrix} \begin{bmatrix} 1 \\ \xi_1 \\ \vdots \\ \xi_{3n}^d \end{bmatrix}^\top \mathrm{d}\,\mu,$$

  it is positive semidefinite. The notation suggests $\mathbf{M}_{2d}(\mathbf{m}_{2d})$ is a (linear) function of the truncated moment vector $\mathbf{m}_{2d}$.

- Second, since the tensor entries are third order moments of the measure, elements of $\mathbf{m}_{2d}$ corresponding to these moments are known when $d \geq 2$, giving rise to the second set of linear equations.

- Third, the fact that μ is supported on $\mathbb{K}$ leads to the last set of linear constraints.

Combined with the fact that the objective function $\mu(\mathbb{K}) = \int_{\mathbb{K}} 1 d\mu = \mathbf{m}_{2d}(1)$, the final relaxation is a semidefinite program. We refer the reader to [95, Section 7] and [72, Section 5.2] for more discussions. Apparently, increasing the relaxation order d yields tighter approximations to the original optimization (3.3). Tang and Shah in [72] showed that for symmetric tensor decomposition, in the undercomplete case and under a soft-orthogonality condition, the smallest semidefinite program in the relaxation hierarchy is tight. Remarkably, Nie in [95] provided detailed convergence analysis of using the Lasserre hierarchy for computing tensor nuclear norms of both symmetric and nonsymmetric tensors over both real and complex fields.

3.4 Proof of Theorem 3.1.1

The proof of Theorem 3.1.1 relies on the construction of a dual polynomial that satisfies the Boundedness and Interpolation Property (3.11). The constructed dual polynomial is also essential to the development of tensor completion and denoising using the atomic norm approach.

3.4.1 Outline of the Proof

First of all, we apply the minimum-energy strategy to construct a candidate dual polynomial q. To show the constructed dual polynomial satisfies the BIP (3.11), we partition $\mathbb{K}$ into the far region (analyzed by Lemma 3.4.4) and the near region. For ease of analyzing the near region, we use an angular parameterization to convert it to the angular near region, which is covered by the vertex region (analyzed by Lemma 3.4.6) and band region (analyzed by Lemma 3.4.7). The proof of Theorem 3.1.1 is completed by combining the far and near regions. We summarize this in Figure 3.1.

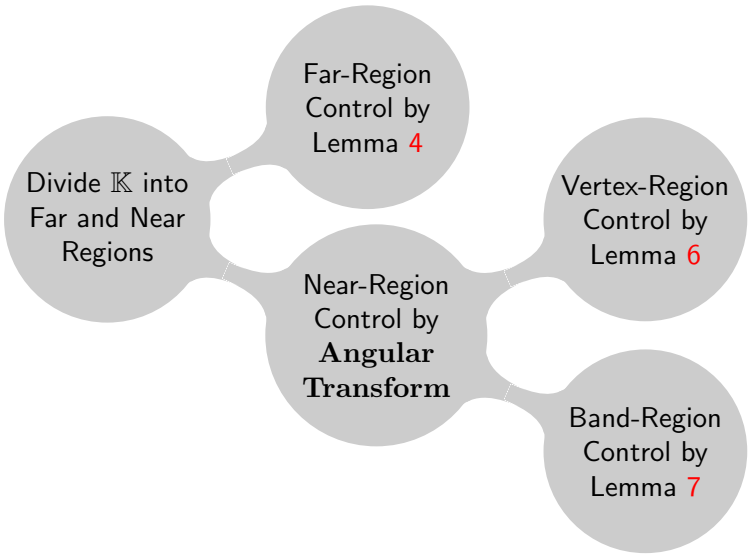


Figure 3.1: An outline of the proof of Theorem 3.1.1.

3.4.2 Minimal-Energy Construction of Pre-certificate

Since the BIP in (3.11) (especially the Boundedness property (3.11b)) is hard to enforce directly, we start from a candidate *dual certificate* or pre-certificate $\boldsymbol{\mathcal{Q}}$ in the subdifferntial set $\partial\|\boldsymbol{\mathcal{T}}\|_*$ defined by (3.16):

$$\begin{aligned}\langle \boldsymbol{\mathcal{Q}}, \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star \rangle &= 1, \ \text{for}\, p \in [r] \\ \langle \boldsymbol{\mathcal{Q}}, \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} \rangle &\leq 1, \ \text{for}(\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathbb{K}\end{aligned}$$

which essentially characterizes the optimal solution set of following optimization

$$\underset{(\mathbf{u},\mathbf{v},\mathbf{w})\in\mathbb{K}}{\text{maximize}}\ \langle \boldsymbol{\mathcal{Q}}, \mathbf{u}\otimes\mathbf{v}\otimes\mathbf{w}\rangle \tag{3.20}$$

Then applying the Karush-Kuhn-Tucker (KKT) conditions to the constrained optimization (3.20), we can further relax the subdifferential conditions (3.16) to a set of linear constraints.

Lemma 3.4.1. *The following conditions are necessary for* (3.16)*:*

$$\begin{aligned}
\sum_{j,k} Q_{ijk}\mathbf{v}_p^\star(j)\mathbf{w}_p^\star(k) &= \mathbf{u}_p^\star(i), \forall i\in[n], \forall p\in[r];\\
\sum_{i,k} Q_{ijk}\mathbf{u}_p^\star(i)\mathbf{w}_p^\star(k) &= \mathbf{v}_p^\star(j), \forall i\in[n], \forall p\in[r];\\
\sum_{i,j} Q_{ijk}\mathbf{u}_p^\star(i)\mathbf{v}_p^\star(j) &= \mathbf{w}_p^\star(k), \forall i\in[n], \forall p\in[r]
\end{aligned}$$

or in tensor notation

$$\begin{aligned}
\boldsymbol{\mathcal{Q}}\times_2\mathbf{v}_p^\star\times_3\mathbf{w}_p^\star &= \mathbf{u}_p^\star, \forall p\in[r];\\
\boldsymbol{\mathcal{Q}}\times_1\mathbf{u}_p^\star\times_3\mathbf{w}_p^\star &= \mathbf{v}_p^\star, \forall p\in[r];\\
\boldsymbol{\mathcal{Q}}\times_1\mathbf{u}_p^\star\times_2\mathbf{v}_p^\star &= \mathbf{w}_p^\star, \forall p\in[r]
\end{aligned} \tag{3.21}$$

where $\{\times_k\}$ *are the* k*-mode tensor-vector product [76] whose definitions are apparent from context.*

The proof of Lemma 3.4.1 is given in Appendix B.1.

Apparently, the subdifferntial conditions (3.16) is necessary for the BIP (3.11), but generally not sufficient, by comparing the second line of (3.16) and the Boundedness Property (3.11b). Indeed, as we argued before, any $\boldsymbol{\mathcal{Q}}$ satisfying the BIP is an interior point of the subdifferential $\partial\|\cdot\|_*(\boldsymbol{\mathcal{T}})$. To satisfy the Boundedness Property (3.11b), we further minimize the energy $\|\boldsymbol{\mathcal{Q}}\|_F^2 = \sum_{ijk} Q_{ijk}^2$ in the hope that this will push $q(\mathbf{u},\mathbf{v},\mathbf{w})$ towards zero such that $\boldsymbol{\mathcal{Q}}$ is an interior point of $\partial\|\cdot\|_*(\boldsymbol{\mathcal{T}})$. Thus, we propose solving the following *minimum-energy* problem to obtain a pre-certificate:

$$\underset{Q}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{\mathcal{Q}}\|_F^2 \ \text{ subject to } (3.21) \tag{3.22}$$

Lemma 3.4.2 (Explicit form of the pre-certificate). *The solution of the least-norm problem (3.22) has the form (normal equation)*

$$\boldsymbol{\mathcal{Q}} = \sum_{p=1}^{r}(\boldsymbol{\alpha}_p^\star\otimes\mathbf{v}_p^\star\otimes\mathbf{w}_p^\star + \mathbf{u}_p^\star\otimes\boldsymbol{\beta}_p^\star\otimes\mathbf{w}_p^\star + \mathbf{u}_p^\star\otimes\mathbf{v}_p^\star\otimes\boldsymbol{\gamma}_p^\star) \tag{3.23}$$

with the unknown coefficients $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ *being chosen such that* $\boldsymbol{\mathcal{Q}}$ *in* (3.23) *satisfies* (3.21)*. So we get an explicit form of a pre-certificate*

$$q(\mathbf{u},\mathbf{v},\mathbf{w}) = \langle \boldsymbol{\mathcal{Q}}, \mathbf{u}\otimes\mathbf{v}\otimes\mathbf{w}\rangle$$
$$= \sum_{p=1}^{r}[\langle \boldsymbol{\alpha}_p^\star, \mathbf{u}\rangle\langle \mathbf{v}_p^\star, \mathbf{v}\rangle\langle \mathbf{w}_p^\star, \mathbf{w}\rangle + \langle \mathbf{u}_p^\star, \mathbf{u}\rangle\langle \boldsymbol{\beta}_p^\star, \mathbf{v}\rangle\langle \mathbf{w}_p^\star, \mathbf{w}\rangle + \langle \mathbf{u}_p^\star, \mathbf{u}\rangle\langle \mathbf{v}_p^\star, \mathbf{v}\rangle\langle \boldsymbol{\gamma}_p^\star, \mathbf{w}\rangle]. \tag{3.24}$$

The proof of Lemma 3.4.2 is given in Appendix B.2.

To obtain some intuition of what these dual-polynomial coefficients $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ would look like, let us assume $\{\mathbf{u}_p^\star\}_{p=1}^r$, $\{\mathbf{v}_p^\star\}_{p=1}^r$, $\{\mathbf{w}_p^\star\}_{p=1}^r$ are almost orthogonal and plug the explicit form of $\boldsymbol{\mathcal{Q}}$ (3.23) into the first equation in (3.21)

$$\boldsymbol{\alpha}_p^\star + \mathbf{u}_p^\star\langle \boldsymbol{\beta}_p^\star, \mathbf{v}_p^\star\rangle + \mathbf{u}_p^\star\langle \boldsymbol{\gamma}_p^\star, \mathbf{w}_p^\star\rangle \approx \mathbf{u}_p^\star. \tag{3.25}$$

Then multiplying $\mathbf{u}_p^{\star\top}$ on both sides gives

$$\langle \boldsymbol{\alpha}_p^\star, \mathbf{u}_p^\star\rangle + \langle \boldsymbol{\beta}_p^\star, \mathbf{v}_p^\star\rangle + \langle \boldsymbol{\gamma}_p^\star, \mathbf{w}_p^\star\rangle \approx 1. \tag{3.26}$$

Finally combining (3.25) and (3.26) together with the symmetry property of (3.23), we get these coefficients $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ are located approximately at $\{\mathbf{u}_p^\star/3, \mathbf{v}_p^\star/3, \mathbf{w}_p^\star/3\}_{p=1}^r$. The accurate description of this phenomenon is given by the following lemma with the proof listed in Appendix B.3.

Lemma 3.4.3 (Control the dual polynomial coefficients). *Under Assumptions II and III together with $r = o(n^2/\kappa(\log n)^2)$, the following estimates are valid for sufficiently large n:*

$$\left\|\mathbf{A} - \frac{1}{3}\mathbf{U}\right\| \le 2\kappa(\log n)\left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}}\right);$$
$$\left\|\mathbf{B} - \frac{1}{3}\mathbf{V}\right\| \le 2\kappa(\log n)\left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}}\right);$$
$$\left\|\mathbf{C} - \frac{1}{3}\mathbf{W}\right\| \le 2\kappa(\log n)\left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}}\right)$$

where

$$\mathbf{A} = \left[\boldsymbol{\alpha}_1^\star, \cdots, \boldsymbol{\alpha}_r^\star\right], \mathbf{U} = \left[\mathbf{u}_1^\star, \cdots, \mathbf{u}_r^\star\right];$$
$$\mathbf{B} = \left[\boldsymbol{\beta}_1^\star, \cdots, \boldsymbol{\beta}_r^\star\right], \mathbf{V} = \left[\mathbf{v}_1^\star, \cdots, \mathbf{v}_r^\star\right];$$
$$\mathbf{C} = \left[\boldsymbol{\gamma}_1^\star, \cdots, \boldsymbol{\gamma}_r^\star\right], \mathbf{W} = \left[\mathbf{w}_1^\star, \cdots, \mathbf{w}_r^\star\right]$$

and the norm $\|\cdot\|$ is the matrix spectral norm.

3.4.3 Far Region

For a parameter $\delta \in (0,1)$, the far region is defined by

$$\mathcal{F}(\delta) := \bigcap_{p=1}^{r} \{(\mathbf{u},\mathbf{v},\mathbf{w}) \in \mathbb{K} : |\langle \mathbf{u}, \mathbf{u}_p^\star\rangle| \le \delta \text{ or } |\langle \mathbf{v}, \mathbf{v}_p^\star\rangle| \le \delta \text{ or } |\langle \mathbf{w}, \mathbf{w}_p^\star\rangle| \le \delta\}, \tag{3.27}$$

which consists of points $(\mathbf{u},\mathbf{v},\mathbf{w})$ in $\mathbb{K}$ that are far away (in the angular sense) from

$$\bar{S}^\star = \{(\pm\mathbf{u}_p^\star, \pm\mathbf{v}_p^\star, \pm\mathbf{w}_p^\star) : p = 1,\dots,r\} \tag{3.28}$$

in at least one coordinate of $(\mathbf{u},\mathbf{v},\mathbf{w})$. For $n = 3$ and $r = 2$, the far region projected onto the unit sphere $\{\mathbf{u} : \|\mathbf{u}\|_2 = 1\}$ is shown in Figure 3.2.

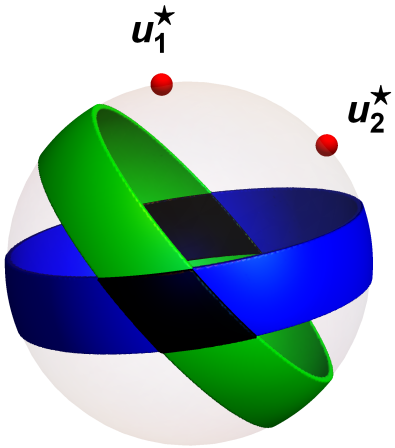


Figure 3.2: Projection of the far region in the $\mathbf{u}$ coordinate. The blue band represents the region $\{\mathbf{u} : |\langle \mathbf{u}, \mathbf{u}_1^\star\rangle| \le \delta\}$ that is far away from $\mathbf{u}_1^\star$, while the green region $\{\mathbf{u} : |\langle \mathbf{u}, \mathbf{u}_2^\star\rangle| \le \delta\}$ is the far-region associated with $\mathbf{u}_2^\star$. The far region is their intersection $\bigcap_{p=1}^{2}\{\mathbf{u} : |\langle \mathbf{u}, \mathbf{u}_p^\star\rangle| \le \delta\}$, consisting of the two black diamonds.

Far-Region Bound. Instead of bounding the dual polynomial q directly, we will bound its absolute value $|q|$. To obtain some intuition of how to bound it, we rewrite the explicit form (3.24) as follows

$$\begin{aligned} & q(\mathbf{u},\mathbf{v},\mathbf{w}) \\ &= \sum_{p=1}^{r}\left[\langle \boldsymbol{\alpha}_p^\star - \frac{1}{3}\mathbf{u}_p^\star, \mathbf{u}\rangle\langle \mathbf{v}_p^\star, \mathbf{v}\rangle\langle \mathbf{w}_p^\star, \mathbf{w}\rangle + \langle \mathbf{u}_p^\star, \mathbf{u}\rangle\langle \boldsymbol{\beta}_p^\star - \frac{1}{3}\mathbf{v}_p^\star, \mathbf{v}\rangle\langle \mathbf{w}_p^\star, \mathbf{w}\rangle + \langle \mathbf{u}_p^\star, \mathbf{u}\rangle\langle \mathbf{v}_p^\star, \mathbf{v}\rangle\langle \boldsymbol{\gamma}_p^\star - \frac{1}{3}\mathbf{w}_p^\star, \mathbf{w}\rangle\right] \qquad (3.29) \\ &+ \sum_{p=1}^{r}\langle \mathbf{u}_p^\star, \mathbf{u}\rangle\langle \mathbf{v}_p^\star, \mathbf{v}\rangle\langle \mathbf{w}_p^\star, \mathbf{w}\rangle. \qquad (3.30) \end{aligned}$$

The main idea is first using the closeness of $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ and $\{\mathbf{u}_p^\star/3, \mathbf{v}_p^\star/3, \mathbf{w}_p^\star/3\}_{p=1}^r$ to bound (3.29) and then using angular-distance between $\mathcal{F}(\delta)$ and $(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star), \forall p$ to bound (3.30).

The accurate argument is made by the following lemma with the proof given in Appendix B.4.

Lemma 3.4.4 (Far-region bound). *Under Assumptions I, II, III, if* $r \ll n^{1.25}$ *and* $r \leq \frac{n}{24\delta c^2}$ *for* $\delta \in (0, \frac{1}{24}]$*, then for sufficiently large* n*, we have* $|q(\mathbf{u}, \mathbf{v}, \mathbf{w})| < 1$ *in* $\mathcal{F}(\delta)$.

3.4.4 Near Region

For the union of the far and near regions to cover the entire region $\mathbb{K}$, we define the near region as

$$\mathcal{N}(\delta) := \mathbb{K} \setminus \mathcal{F}(\delta) = \bigcup_{p=1}^{r} \{(\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathbb{K} : |\langle \mathbf{u}_p^\star, \mathbf{u} \rangle| \geq \delta, |\langle \mathbf{v}_p^\star, \mathbf{v} \rangle| \geq \delta, |\langle \mathbf{w}_p^\star, \mathbf{w} \rangle| \geq \delta\} \tag{3.31}$$

using *De Morgan's Law*. One can also treat the whole near region as a union of all individual ones

$$\mathcal{N}(\delta) = \bigcup_{p=1}^{r} \mathcal{N}_p(\delta)$$

with each individual near region defined by

$$\mathcal{N}_p(\delta) := \{(\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathbb{K} : |\langle \mathbf{u}_p^\star, \mathbf{u} \rangle| \geq \delta, |\langle \mathbf{v}_p^\star, \mathbf{v} \rangle| \geq \delta, |\langle \mathbf{w}_p^\star, \mathbf{w} \rangle| \geq \delta\} \tag{3.32}$$

which is composed of all the points that is closed to at least one point in $\bar{S}^\star$ in all coordinate of $(\mathbf{u}, \mathbf{v}, \mathbf{w})$. For $n = 3$, $r = 2$, we plot the near region $\mathcal{N}_1(\delta)$ projected onto the sphere $\{\mathbf{u} : \|\mathbf{u}\|_2 = 1\}$ in Figure 3.3.

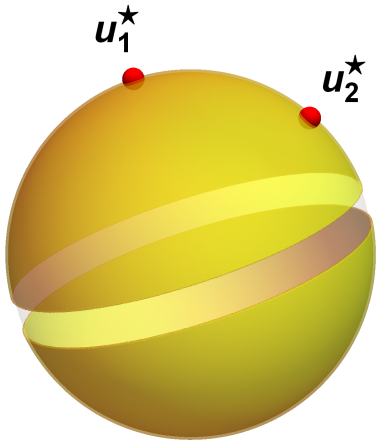


Figure 3.3: The two yellow spherical caps form the near region $\mathcal{N}_1(\delta)$ around the point $(\mathbf{u}_1^\star, \mathbf{v}_1^\star, \mathbf{w}_1^\star)$ projected onto the $\mathbf{u}$ coordinates. $\mathcal{N}_2(\delta)$, which is not shown here, consists of another two spherical caps. The union of $\mathcal{N}_1(\delta), \mathcal{N}_2(\delta)$ and the far region $\mathcal{F}(d)$ shown in Figure 3.2 will cover the entire sphere $\{\mathbf{u} : \|\mathbf{u}\| = 1\}$.

In order to show the dual polynomial satisfying the BIP in the entire near region $\mathcal{N}(\delta)$, we use the "Divide-and-conquer" idea to bound the dual polynomial in each individual near region $\mathcal{N}_p(\delta)$ for $p \in [r]$. The main technique used to control each individual near region is applying angular parameterization to each individual near region.

3.4.5 Angular Parameterization

Angular Parametrization of Near Region. As the domain $\mathbb{K}$ is essentially a direct product of spheres, we re-parameterize each individual near region $\mathcal{N}_p(\delta)$ in the angular sense. Without loss of generality, let us consider $p = 1$. Pick $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathbb{K}$ such that $\mathbf{x} \perp \mathbf{u}_1^\star, \mathbf{y} \perp \mathbf{v}_1^\star, \mathbf{z} \perp \mathbf{w}_1^\star$ and consider the parameterized points

$$(\mathbf{u}(\theta_1), \mathbf{v}(\theta_2), \mathbf{w}(\theta_3)) \in \mathbb{K}$$

with

$$\begin{aligned} \mathbf{u}(\theta_1) &= \mathbf{u}_1^\star \cos(\theta_1) \ + \mathbf{x} \sin(\theta_1), \\ \mathbf{v}(\theta_2) &= \mathbf{v}_1^\star \cos(\theta_2) \ + \mathbf{y} \sin(\theta_2), \\ \mathbf{w}(\theta_3) &= \mathbf{w}_1^\star \cos(\theta_3) + \mathbf{z} \sin(\theta_3). \end{aligned} \tag{3.33}$$

When θ_1 ranges from 0 to π, $\mathbf{u}(\theta_1)$ traces out a 2D semi-circle that starts at $\mathbf{u}_1^\star$, passes through $\mathbf{x}$, and finally reaches $-\mathbf{u}_1^\star$; while for a fixed $\theta_1 \in [0, \pi]$, the set $\bigcup_{\mathbf{x} \perp \mathbf{u}_1^\star} \{\mathbf{u}(\theta_1)\}$ parameterizes all the points on $\mathbb{S}^{n-1}$ having an angle of θ_1 with $\mathbf{u}_1^\star$. The same properties hold for $\mathbf{v}(\theta_2)$ and $\mathbf{w}(\theta_3)$. This parametrization projected onto the $\mathbf{u}$ coordinate is shown in Figure 3.4.

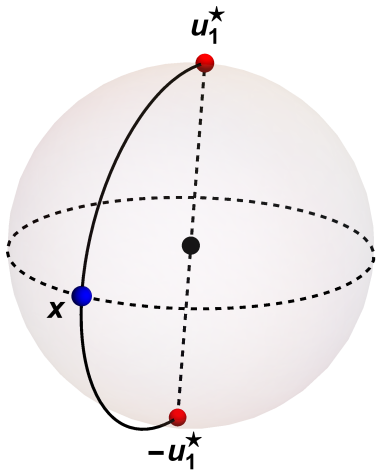


Figure 3.4: Parameterization of points on the unit sphere for $\mathbf{u}$.

In fact, using this angular parametrization, the individual near region $\mathcal{N}_1(\delta)$ in (3.32) can be expressed as

$$\mathcal{N}_1(\delta) = \bigcup_{(\mathbf{x},\mathbf{y},\mathbf{z}):\mathbf{x}\perp\mathbf{u}_1^\star,\mathbf{y}\perp\mathbf{v}_1^\star,\mathbf{z}\perp\mathbf{w}_1^\star} \{(\mathbf{u}(\theta_1), \mathbf{v}(\theta_2), \mathbf{w}(\theta_3)) : |\cos(\theta_i)| \geq \delta, \theta_i \in [0, \pi], i = 1, 2, 3\}. \tag{3.34}$$

Proposition 3.4.1 (Angular near region). *For any $\delta \in (0, 1)$, the near region $\mathcal{N}_1(\delta)$ is contained in the following set*

$$\mathcal{N}_1(\delta) \subset \bigcup_{(\mathbf{x},\mathbf{y},\mathbf{z}):\mathbf{x}\perp\mathbf{u}_1^\star,\mathbf{y}\perp\mathbf{v}_1^\star,\mathbf{z}\perp\mathbf{w}_1^\star} \{(\mathbf{u}(\theta_1), \mathbf{v}(\theta_2), \mathbf{w}(\theta_3)) : (\theta_1, \theta_2, \theta_3) \in \mathbb{N}(\delta)\} \tag{3.35}$$

with the angular near region $\mathbb{N}(\delta)$ *defined by*

$$\mathbb{N}(\delta) := \left\{ (\theta_1, \theta_2, \theta_3) : \theta_i \in \left[0, \frac{\pi}{2} - \delta\right] \cup \left[\frac{\pi}{2} + \delta, \pi\right], i = 1, 2, 3 \right\}. \tag{3.36}$$

Proof. Since the function $|\cos(\theta)|$ is symmetric at $\frac{\pi}{2}$ on the interval $[0, \pi]$ and is decreasing on $[0, \pi/2]$, we know that $\{\theta : |\cos(\theta)| \geq \delta\} \cap [0, \pi] = [0, \arccos(\delta)] \cup [\pi - \arccos(\delta), \pi]$. Note that $\arccos(\delta) = \frac{\pi}{2} - \arcsin(\delta)$ and $\delta < \arcsin(\delta)$, so we get $\{\theta : |\cos(\theta)| \geq \delta\} \cap [0, \pi] \subset [0, \frac{\pi}{2} - \delta] \cup [\frac{\pi}{2} + \delta, \pi]$. The inclusion (3.35) follows from (3.34) immediately. □

The angular near region $\mathbb{N}(\delta)$ contains total eight cubes with side length $\frac{\pi}{2} - \delta$, located at the eight corners of the cube $[0, \pi] \times [0, \pi] \times [0, \pi]$. Moreover, one can see that the smaller the parameter δ is, the larger the angular near region $\mathbb{N}(\delta)$ will be. In particular, when δ approaches to zero, the angular near region $\mathbb{N}(\delta)$ becomes the whole cube $\mathbb{N}(0) = [0, \pi] \times [0, \pi] \times [0, \pi]$. The angular near region $\mathbb{N}(\delta)$ is plotted in Figure 3.5.

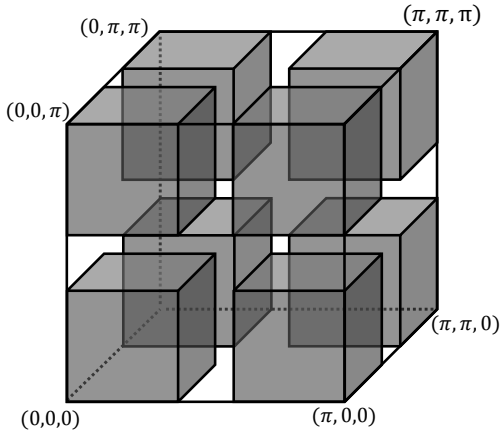


Figure 3.5: The eight gray cubes of side-length $\pi/2 - \delta$ at the corners form the angular near region $\mathbb{N}(\delta)$.

Angular Parametrization of Dual Polynomial. Evaluating the dual polynomial $q(\mathbf{u}, \mathbf{v}, \mathbf{w})$ at $(\mathbf{u}(\theta_1), \mathbf{v}(\theta_2), \mathbf{w}(\theta_3))$ in (3.33), we get the angular dual polynomial $F(\theta_1, \theta_2, \theta_3) := q(\mathbf{u}(\theta_1), \mathbf{v}(\theta_2), \mathbf{w}(\theta_3))$ as

$$\begin{aligned} F(\theta_1, \theta_2, \theta_3) =& q(\mathbf{u}_1^\star, \mathbf{v}_1^\star, \mathbf{w}_1^\star) \cos(\theta_1) \cos(\theta_2) \cos(\theta_3) \\ &+ q(\mathbf{u}_1^\star, \mathbf{v}_1^\star, \mathbf{z}) \cos(\theta_1) \cos(\theta_2) \sin(\theta_3) \\ &+ q(\mathbf{u}_1^\star, \mathbf{y}, \mathbf{w}_1^\star) \cos(\theta_1) \sin(\theta_2) \cos(\theta_3) \\ &+ q(\mathbf{x}, \mathbf{v}_1^\star, \mathbf{w}_1^\star) \sin(\theta_1) \cos(\theta_2) \cos(\theta_3) \\ &+ q(\mathbf{u}_1^\star, \mathbf{y}, \mathbf{z}) \cos(\theta_1) \sin(\theta_2) \sin(\theta_3) \\ &+ q(\mathbf{x}, \mathbf{v}_1^\star, \mathbf{z}) \sin(\theta_1) \cos(\theta_2) \sin(\theta_3) \\ &+ q(\mathbf{x}, \mathbf{y}, \mathbf{w}_1^\star) \sin(\theta_1) \sin(\theta_2) \cos(\theta_3) \\ &+ q(\mathbf{x}, \mathbf{y}, \mathbf{z}) \sin(\theta_1) \sin(\theta_2) \sin(\theta_3). \end{aligned}$$

Among these 8 terms, the first term is $\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)$ since $q(\mathbf{u}_1^\star, \mathbf{v}_1^\star, \mathbf{w}_1^\star) = 1$. The next three terms involving one sine function are zero as, for example,

$$q(\mathbf{u}_1^\star, \mathbf{v}_1^\star, \mathbf{z}) = \boldsymbol{\mathcal{Q}} \times_1 \mathbf{u}_1^\star \times_2 \mathbf{v}_1^\star \times_3 \mathbf{z} = \mathbf{w}_1^\star \times_3 \mathbf{z} = {\mathbf{w}_1^\star}^\top \mathbf{z} = 0,$$

where we have used $Q \times_1 \mathbf{u}_1^\star \times_2 \mathbf{v}_1^\star = \mathbf{w}_1^\star$ and the third equality of (3.21). Hence, we get a more concise form of F:

$$\begin{aligned} F(\theta_1, \theta_2, \theta_3) = \cos(\theta_1)\cos(\theta_2)\cos(\theta_3) &+ q(\mathbf{u}_1^\star, \mathbf{y}, \mathbf{z})\cos(\theta_1)\sin(\theta_2)\sin(\theta_3) \\ &+ q(\mathbf{x}, \mathbf{v}_1^\star, \mathbf{z})\sin(\theta_1)\cos(\theta_2)\sin(\theta_3) \\ &+ q(\mathbf{x}, \mathbf{y}, \mathbf{w}_1^\star)\sin(\theta_1)\sin(\theta_2)\cos(\theta_3) \\ &+ q(\mathbf{x}, \mathbf{y}, \mathbf{z})\sin(\theta_1)\sin(\theta_2)\sin(\theta_3). \end{aligned} \tag{3.37}$$

By further bounding the other quantities $q(\mathbf{u}_1^\star, \mathbf{y}, \mathbf{z})$, $q(\mathbf{x}, \mathbf{v}_1^\star, \mathbf{z})$, $q(\mathbf{x}, \mathbf{y}, \mathbf{w}_1^\star)$ and $q(\mathbf{x}, \mathbf{y}, \mathbf{z})$, we get the following lemma to uniformly upper-bound $F(\theta_1, \theta_2, \theta_3)$ with the proof given in Appendix B.5.

Lemma 3.4.5 (Uniform upper bound of F). *Under Assumptions I, II, III, if $r \le n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$, then for sufficiently large n, we have*

$$|F(\theta_1, \theta_2, \theta_3)| \le |\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)| + |\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)| + \frac{4}{3}\tau(\log n)n^{-r_c}. \tag{3.38}$$

3.4.6 Near-Region Bound

Angular Boundedness and Interpolation Property. By Proposition 3.4.1, a sufficient condition for the BIP (3.11) to hold in the individual near region $\mathcal{N}_1(\delta)$, is the following Angular Boundedness and Interpolation Property (Angular-BIP):

$$F(\theta_1, \theta_2, \theta_3) = 1 \text{ in } \mathbb{S}^\star \ \text{(Angular Interpolation)} \tag{3.39a}$$
$$F(\theta_1, \theta_2, \theta_3) < 1 \text{ in } \mathbb{N}(\delta) \setminus \mathbb{S}^\star \ \text{(Angular Bounedness)} \tag{3.39b}$$

with $\mathbb{S}^\star := \{(0,0,0), (0,\pi,\pi), (\pi,0,\pi), (\pi,\pi,0)\}$ such that $\mathbf{u}(\theta_1) \otimes \mathbf{v}(\theta_2) \otimes \mathbf{w}(\theta_3) = \mathbf{u}_1^\star \otimes \mathbf{v}_1^\star \otimes \mathbf{w}_1^\star$ for any $(\theta_1, \theta_2, \theta_3) \in \mathbb{S}^\star$.

Similar as before, the Angular Interpolation property (3.39a) is a consequence of the construction process. In the rest of the chapter, we will focus on showing the Angular Boundedness property (3.39b). Specifically, we will control the angular dual polynomial F in both the vertex region and band region and then show their union covers the angular near region $\mathbb{N}(\delta)$.

Vertex Region. The vertex region, denoted by $\mathbb{N}_v(\delta_v)$, is defined as the union of the eight small cubes all with side length δ_v in 8 corners of the cube $[0, \pi]^3$. We plot the vertex region $\mathbb{N}_v(\delta_v)$ in Figure 3.6. Comparing with the definition of the angular near region $\mathbb{N}(\cdot)$, the vertex region is also an angular near region but with a different parameter:

$$\mathbb{N}_v(\delta_v) = \mathbb{N}(\frac{\pi}{2} - \delta_v). \tag{3.40}$$

Without loss of generality, we can always assume the vertex region $\mathbb{N}_v(\delta_v)$ is included in the angular near region $\mathbb{N}(\delta)$; otherwise, we only need to show the Angular-BIP holds in $\mathbb{N}_v(\delta_v)$. This assumption together with (3.40) implies

$$\delta_v \leq \frac{\pi}{2} - \delta. \tag{3.41}$$

Note that $\pi/2 - \delta$ is the side length of the corner-cubes in $\mathbb{N}(\delta)$.

Vertex-Region Bound. To control the angular dual polynomial F in the vertex region $\mathbb{N}_v(\delta_v)$, we further classify the eight small cubes in $\mathbb{N}_v(\delta_v)$ into two groups depending on if their vertices are in $\mathbb{S}^\star$ or not.

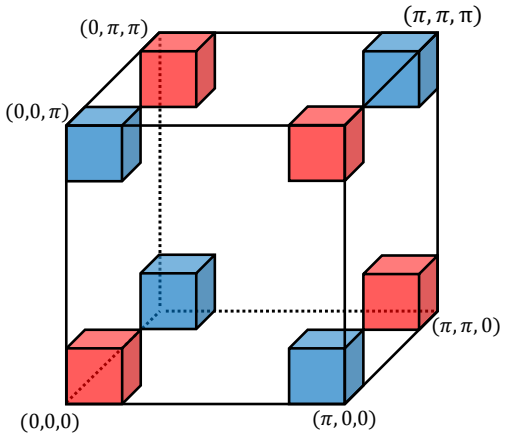


Figure 3.6: The eight colored cubes of size $\delta_v \times \delta_v \times \delta_v$ form the vertex region $\mathbb{N}_v(\delta_v)$: the red ones are corresponding to the vertexes in $\mathbb{S}^\star$ while the blue ones are corresponding to other vertexes in the cube. Note that these colored corner-cubes are possibly much smaller than those gray ones in Figure 3.5, whose side length is $\pi/2 - \delta$.

Lemma 3.4.6 (Vertex-region bound). *Under Assumptions I, II, III, if* $r \ll n^{1.25}$*, then for any* $\xi_i \in \left(-\frac{\sqrt{2}-1}{3}, \frac{\sqrt{2}-1}{3}\right)$*, we have*

$$F(\theta_1 + \xi_1, \theta_2 + \xi_2, \theta_3 + \xi_3) \leq 1 \tag{3.42}$$

for $(\theta_1, \theta_2, \theta_3) \in \{(0,0,0), (0,\pi,\pi), (\pi,0,\pi), (\pi,\pi,0)\}$ *and*

$$F(\theta_1 + \xi_1, \theta_2 + \xi_2, \theta_3 + \xi_3) < 0 \tag{3.43}$$

for $(\theta_1, \theta_2, \theta_3) \in \{(\pi,\pi,\pi), (\pi,0,0), (0,\pi,0), (0,0,\pi)\}$*. Here, equality in* (3.42) *holds only if* $\xi_1 = \xi_2 = \xi_3 = 0$*.*

The proof of Lemma 3.4.6 is in Appendix B.6.

Remark 3.4.1. Lemma 3.4.6 proves the Angular-BIP (3.39) holds in the vertex region $\mathbb{N}_v(\delta_v)$ with $\delta_v = \frac{\sqrt{2}-1}{3}$:

$$F(\theta_1, \theta_2, \theta_3) = 1 \text{ in } \mathbb{S}^\star$$
$$F(\theta_1, \theta_2, \theta_3) < 1 \text{ in } \mathbb{N}_v(\delta_v) \setminus \mathbb{S}^\star$$

Band Region. The band region is introduced to cover the remaining region $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$. Invoking the definitions of the angular near region (3.36) and the vertex region (3.40):

$$\mathbb{N}(\delta) = \left\{(\theta_1, \theta_2, \theta_3) : \theta_i \in \left[0, \frac{\pi}{2} - \delta\right] \cup \left[\frac{\pi}{2} + \delta, \pi\right]\right\}$$
$$\mathbb{N}_v(\delta_v) = \{(\theta_1, \theta_2, \theta_3) : \theta_i \in [0, \delta_v] \cup [\pi - \delta_v, \pi]\}$$

we have

$$\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v) = \left\{(\theta_1, \theta_2, \theta_3) : \theta_i \in \left(\delta_v, \frac{\pi}{2} - \delta\right) \cup \left(\frac{\pi}{2} + \delta, \pi - \delta_v\right)\right\} \cap \mathbb{N}(\delta), \tag{3.44}$$

which is nonempty since $\delta_v \leq \pi/2 - \delta$ by the assumption (3.41). We plot the remaining region $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$ projected onto the (θ_1, θ_2)-coordinates in Figure 3.7.

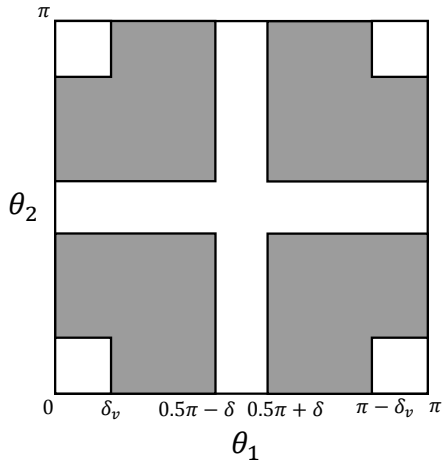


Figure 3.7: The remaining region $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$ projected onto the (θ_1, θ_2)-coordinates.

To let the band region cover $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$, we define it as

$$\mathbb{N}_b(\delta_b) := \left\{(\theta_1, \theta_2, \theta_3) : \theta_i \in \left(\delta_b, \frac{\pi}{2} - \delta_b\right) \cup \left(\frac{\pi}{2} + \delta_b, \pi - \delta_b\right), i = 1, 2, 3\right\}. \tag{3.45}$$

We plot the band region $\mathbb{N}_b(\delta_b)$ projected onto the (θ_1, θ_2)-coordinates in Figure 3.8.

Remark 3.4.2. From (3.44) and (3.45), we have $\mathbb{N}_b(\delta_b)$ covers $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$ if $\delta_b \leq \min\{\delta_v, \delta\}$, or equivalently,

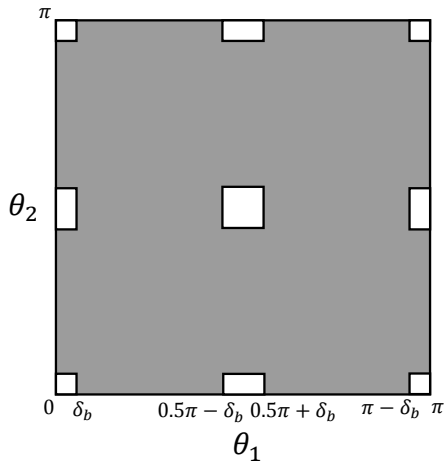


Figure 3.8: The band region $\mathbb{N}_b(\delta_b)$ projected onto the (θ_1, θ_2)-coordinates. Clearly, when $\delta_b \leq \min\{\delta_v, \delta\}$, the band region $\mathbb{N}_b(\delta_b)$ covers the remaining region $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$, as plotted in Figure 3.7.

$$\mathbb{N}(\delta) \subset \mathbb{N}_b(\delta_b) \cup \mathbb{N}_v(\delta_v), \quad \text{if } \delta_b \leq \min\{\delta_v, \delta\}. \tag{3.46}$$

Band-Region Bound. We start with the uniform upper-bound in Lemma 3.4.5:

$$\begin{aligned}
|F(\theta_1, \theta_2, \theta_3)| \leq & |\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)| + |\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)| + \frac{4}{3}\tau(\log n)n^{-r_c} \\
\leq & \frac{1}{3}(|\cos(\theta_1)|^3 + |\cos(\theta_2)|^3 + |\cos(\theta_3)|^3) + \frac{1}{3}(|\sin(\theta_1)|^3 + |\sin(\theta_2)|^3 + |\sin(\theta_3)|^3) + \frac{4}{3}\tau(\log n)n^{-r_c} \\
\leq & \frac{1}{3}(|\cos(\theta_i)|^3 + |\sin(\theta_i)|^3) + \frac{2}{3} + \frac{4}{3}\tau(\log n)n^{-r_c}, \ \forall i \in \{1, 2, 3\}
\end{aligned} \tag{3.47}$$

where the first inequality follows from (3.38) in Lemma 3.4.5 (under Assumptions I-III and $r \leq n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$), the second inequality follows from the inequality of arithmetic and geometric means, and the last one is a consequence of $|\sin(\theta)|^3 + |\cos(\theta)|^3 \leq 1$. So, $|F(\theta_1, \theta_2, \theta_3)| < 1$ in $\mathbb{N}_b(\delta_b)$ if

$$|\cos(\theta_i)|^3 + |\sin(\theta_i)|^3 < 1 - 4\tau(\log n)n^{-r_c} \tag{3.48}$$

for *some* $i \in \{1, 2, 3\}$. The final result is summarized in the following lemma, with the proof listed in Appendix B.7.

Lemma 3.4.7 (Band-region bound). *Under Assumptions I, II, III, if* $r \leq n^{1.25-1.5r_c}$ *with* $r_c \in (0, \frac{1}{6})$*, then for sufficiently large* n*, we have* $|F(\theta_1, \theta_2, \theta_3)| < 1$ *in* $\mathbb{N}_b(\delta_b)$ *for* $\delta_b = \sqrt{\frac{80\tau(\log n)}{3}}n^{-0.5r_c}$.

Combine the Vertex and Band Regions. Finally the Angular-BIP (3.39) follows from Lemma 3.4.6 and Lemma 3.4.7 if the union of the vertex region $\mathbb{N}_v(\delta_v)$ and the band region $\mathbb{N}_b(\delta_b)$ covers the angular near region $\mathbb{N}(\delta)$:

$$\mathbb{N}(\delta) \subset \mathbb{N}_v(\delta_v) \cup \mathbb{N}_b(\delta_b).$$

From (3.46), this happens when

$$\delta_b \leq \min\{\delta, \delta_v\},$$

which is equivalent to

$$\delta_b \leq \delta, \tag{3.49}$$

since

$$\delta_b = \sqrt{\frac{80\tau(\log n)}{3}} n^{-0.5r_c} \ll \frac{\sqrt{2}-1}{3} = \delta_v.$$

Then by Proposition 3.4.1, q satisfies the BIP in $\mathcal{N}_1(\delta)$. Similar results apply to all individual near region $\mathcal{N}_p(\delta)$, for $p \in [r]$. Therefore we claim the BIP holds in the whole near region $\mathcal{N}(\delta) = \bigcup_{p=1}^{r} \mathcal{N}_p(\delta)$.

Lemma 3.4.8 (Near-region bound). *Under Assumptions I, II, III, if $r \leq n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$, then for sufficiently large n, the dual polynomial q satisfies the BIP in $\mathcal{N}(\delta)$ for any $\delta \geq \delta_b$.*

3.4.7 Combine the Far and Near Regions

Combining Lemma 3.4.4 (for far region) and Lemma 3.4.8 (for near region), we conclude that the BIP holds in the whole domain $\mathbb{K}$ if Assumptions I, II, III are satisfied and

$$r \leq \frac{n}{24\delta c^2} \text{ for } \delta \in \left[\delta_b, \frac{1}{24}\right], \tag{3.50}$$

$$r \leq n^{1.25-1.5r_c} \text{ for } r_c \in \left(0, \frac{1}{6}\right). \tag{3.51}$$

Then letting $\delta = \delta_b$ (to maximize r) and $r_c = \frac{1}{8}$, the requirements on r ((3.50) and (3.51)) are reduced to the desired bound (3.7):

$$r \leq \frac{n^{17/16}}{32c^2\sqrt{15\tau(\log n)}}.$$

The proof of Theorem 3.1.1 is completed.

□

3.5 Numerical Experiments

In this section, some numerical results are presented to test the performance of the proposed computational methods. In the first experiment, we examine the phase transition curves of the rate of success for three algorithms:

i) ADMM implementation of (3.18) with “Good Initialization" (ADMM-G), ii) ADMM with random initialization (ADMM-R) and iii) the Lasserre hierarchy relaxation of order $d = 2$ (SOS-2). ADMM with “Good Initialization" uses the output of the power method developed in [71] as initialization.

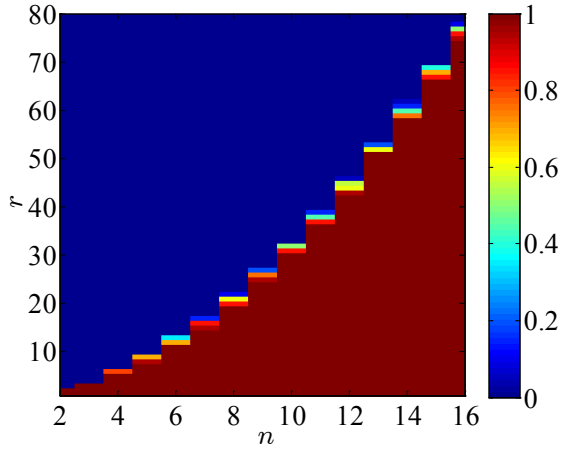
The phase transition curves are plotted in Figure 3.9. In preparing this figure, the r tensor factors $\{(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)\}_{p=1}^r$ were generated following i.i.d. Gaussian distribution, and then each $\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star$ was normalized to have a unit norm. We set the coefficients $\lambda_p^\star = (1+\varepsilon_p^2)/2$, where ε_p is chosen from the standard normal distribution, to ensure a minimal coefficient of at least $1/2$. With the generated ground-truth factors $\{(\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star)\}_{p=1}^r$ and coefficients $\{\lambda_p\}_{p=1}^r$, we generated the tensor $\boldsymbol{\mathcal{T}} = \sum_{p=1}^r \lambda_p^\star \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star$.

To test our theory, we varied the dimension n and factor-number r. For each fixed (r, n) pair, 20 instances of such tensor were generated. We then ran the three algorithms for each instance and declared success if i) the recovered truncated moment vector is within 10^{-3} distance of the true moment vector for the the Lasserre hierarchy relaxation method, and ii) the recovered tensor factors are within 10^{-3} distance to the true tensor factors. We used the moment vector criteria for the Lasserre hierarchy method because one cannot identify more than n tensor factors for the $d = 2$ relaxation. Also, considering the high computational complexity of the Lasserre hierarchy method when n is large, we only set n range from 2 to 8. The rate of success for each algorithm is the percentage of successful instances.

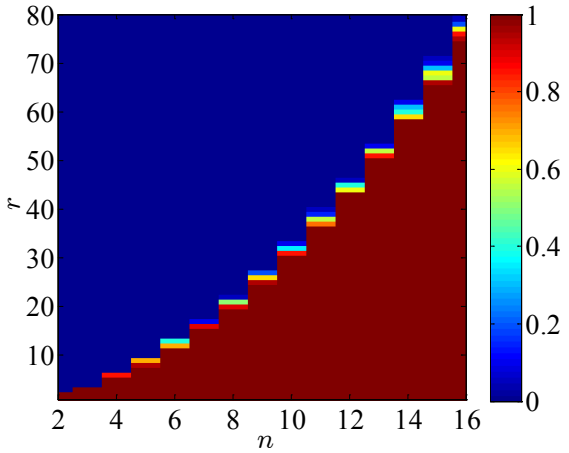
From Figure 3.9, we observe that the Lasserre hierarchy relaxation with $d = 2$ is unable to identify more than n factors. The ADMM method works for r much larger than n. In addition, random initialization does not degrade the performance compared with “Good Initialization".

3.6 Conclusion

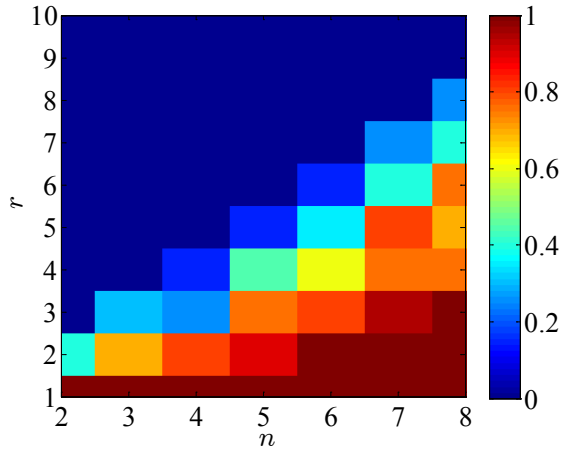
By explicitly constructing a *dual certificate*, we derive conditions for a tensor decomposition to achieve the tensor nuclear norm. This implies that the infinite dimensional measure optimization, which defines the tensor nuclear norm, is able to recover the decomposition under an incoherent condition and two other mild conditions. Computational methods based on low-rank factorization and the Lasserre hierarchy relaxation are used to solve the measure optimization. Numerical experiments show that the nonlinear programming approach has superior performance. Future work will analyze the observed good performance of the nonlinear programming formulation.



(a) ADMM-G



(b) ADMM-R



(c) SOS-2

Figure 3.9: Rate of success for tensor decomposition using ADMM-G, ADMM-R and SOS-2.

CHAPTER 4

THE NONCONVEX GEOMETRY OF LOW-RANK MATRIX OPTIMIZATION

This work[4] considers two popular minimization problems: (i) the minimization of a general convex function $f(\mathbf{X})$ with the domain being positive semi-definite matrices; (ii) the minimization of a general convex function $f(\mathbf{X})$ regularized by the matrix nuclear norm $\|\mathbf{X}\|_*$ with the domain being general matrices. Despite their optimal statistical performance in the literature, these two optimization problems have a high computational complexity even when solved using tailored fast convex solvers. To develop faster and more scalable algorithms, we follow the proposal of Burer and Monteiro to factor the low-rank variable $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$ (for semi-definite matrices) or $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ (for general matrices) and also replace the nuclear norm $\|\mathbf{X}\|_*$ with $(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)/2$. In spite of the non-convexity of the resulting factored formulations, we prove that each critical point either corresponds to the global optimum of the original convex problems or is a strict saddle where the Hessian matrix has a strictly negative eigenvalue. Such a nice geometric structure of the factored formulations allows many local search algorithms to find a global optimizer even with random initializations.

4.1 Introduction

Nonconvex reformulations of convex optimization problems have received a surge of renewed interest for efficiency and scalability reasons [4, 97–112]. Compared with the convex formulations, the non-convex ones typically involve many fewer variables, allowing them to scale to scenarios with millions of variables. Besides, simple algorithms [97,113,114] applied to the non-convex formulations have surprisingly good performance in practice. However, a complete understanding of this phenomenon, particularly the geometrical structures of these non-convex optimization problems, is still an active research area. Unlike the simple geometry of convex optimization problems where local minimizers are also global ones, the landscapes of general non-convex functions can become extremely complicated. Fortunately, for a range of convex optimization problems, particularly for matrix completion and sensing problems, the corresponding non-convex reformulations have nice geometric structures that allow local-search algorithms to converge to global optimality [93, 97, 100, 101, 104, 113, 114] .

We extend this line of investigation by working with a general convex function $f(\mathbf{X})$ and considering the following two popular optimization problems:

$$\text{For symmetric case: } \underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}}\ f(\mathbf{X}) \ \text{ subject to } \ \mathbf{X}\succeq 0 \tag{$\mathcal{P}_0$}$$

$$\text{For nonsymmetric case: } \underset{\mathbf{X}\in\mathbb{R}^{n\times m}}{\text{minimize}}\ f(\mathbf{X}) + \lambda\|\mathbf{X}\|_* \ \text{ where } \lambda > 0 \tag{$\mathcal{P}_1$}$$

[4] This is a joint work with Zhihui Zhu and Gongguo Tang [6].

For these two problems, even fast first-order methods, such as the projected gradient descent algorithm [115], require performing an expensive eigenvalue decomposition or singular value decomposition in each iteration. These expensive operations form the major computational bottleneck and prevent them from scaling to scenarios with millions of variables, a typical situation in a diverse range of applications, including quantum state tomography [116], user preferences prediction [117], and pairwise distances estimation in sensor localization [118].

4.1.1 Our Approach: Burer-Monteiro Style Parameterization

As we have seen, the extremely large dimension of the optimization variable $\mathbf{X}$ and the accordingly expensive eigenvalue or singular value decompositions on X form the major computational bottleneck of the convex optimization algorithms. An immediate question might be "Is there a way to directly reduce the dimension of the optimization variable $\mathbf{X}$ and meanwhile avoid performing the expensive eigenvalue or singular value decompositions?"

This question can be answered when the original optimization problems $(\mathcal{P}_0)$- $(\mathcal{P}_1)$ admit a low-rank solution $\mathbf{X}^\star$ with $\operatorname{rank}(\mathbf{X}^\star) = r^\star \ll \min\{n, m\}$. Then we can follow the proposal of Burer and Monteiro [119] to parameterize the low-rank variable as $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$ for $(\mathcal{P}_0)$ or $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ for $(\mathcal{P}_1)$, where $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{V} \in \mathbb{R}^{m\times r}$ with $r \geq r^\star$. Moreover, since $\|\mathbf{X}\|_* = \operatorname{minimize}_{\mathbf{X}=\mathbf{U}\mathbf{V}^\top}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)/2$, we obtain the following non-convex re-parameterizations of $(\mathcal{P}_0)$- $(\mathcal{P}_1)$:

$$\text{For symmetric case:} \quad \underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\operatorname{minimize}}\, g(\mathbf{U}) = f(\mathbf{U}\mathbf{U}^\top) \tag{$\mathcal{F}_0$}$$

$$\text{For nonsymmetric case:} \quad \underset{\mathbf{U}\in\mathbb{R}^{n\times r}, V\in\mathbb{R}^{m\times r}}{\operatorname{minimize}}\, g(\mathbf{U},\mathbf{V}) = f(\mathbf{U}\mathbf{V}^\top) + \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right) \tag{$\mathcal{F}_1$}$$

Since $r \ll \{p, q\}$, the resulting factored problems $(\mathcal{F}_0)$- $(\mathcal{F}_1)$ involve many fewer variables. Moreover, because the positive semi-definite constraint is removed from $(\mathcal{P}_0)$ and the nuclear norm $\|\mathbf{X}\|_*$ in $(\mathcal{P}_1)$ is replaced by $(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)/2$, there is no need to perform an eigenvalue (or a singular value) decomposition in solving the factored problems.

The past two years have seen renewed interest in the Burer-Monteiro factorization for solving low-rank matrix optimization problems [100–104, 120]. With technical innovations in analyzing the non-convex landscape of the factored objective function, several recent works have shown that with an exact parameterization (i.e., $r = r^\star$) the resulting factored reformulation has no spurious local minima or degenerate saddle points [93, 100, 101, 104]. An important implication is that local-search algorithms such as gradient descent and its variants can converge to the global optima with even random initialization [97, 113, 114].

We generalize this line of work by assuming a general objective function $f(\mathbf{X})$ in $(\mathcal{P}_0)$- $(\mathcal{P}_1)$, not necessarily coming from a matrix inverse problem. This generality allows us to view the resulting factored problems $(\mathcal{F}_0)$- $(\mathcal{F}_1)$ as a way to solve the original convex optimization problems to the global optimum, rather than a new modeling

method. This perspective, also taken by Burer and Monteiro in their original work [119], frees us from rederiving the statistical performances of the resulting factored optimization problems. Instead, the statistical performances of the resulting factored optimization problems inherit from that of the original convex optimization problems, whose statistical performance can be analyzed using a suite of powerful convex analysis techniques, which have accumulated from several decades of research. For example, the original convex optimization problems $(\mathcal{P}_0)$- $(\mathcal{P}_1)$ have information-theoretically optimal sampling complexity [121], achieve minimax denoising rate [122] and satisfy tight oracle inequalities [123]. Therefore, the statistical performances of the factored optimization problems $(\mathcal{F}_0)$- $(\mathcal{F}_1)$ share the same theoretical bounds as those of the original convex optimization problems $(\mathcal{P}_0)$- $(\mathcal{P}_1)$, as long as we can show that the two problems are equivalent.

In spite of their optimal statistical performance [121–124], the original convex optimization problems cannot be scaled to solve the practical problems that originally motivate their development even with specialized first-order algorithms. This was realized since the advent of this field where the low-rank factorization method was proposed as an alternative to convex solvers [119]. When coupled with stochastic gradient descent, low-rank factorization leads to state-of-the-art performance in practical matrix recovery problems [93, 100–102, 104]. Therefore, our general analysis technique also sheds light on the connection between the geometries of the original convex programs and their non-convex reformulations.

Although the Burer-Monteiro parameterization tremendously reduces the number of optimization variables from n^2 to nr (or nm to $(n+m)r$) when r is very small, the intrinsic bi-linearity makes the factored objective functions non-convex and introduces additional critical points that are not global optima of the factored optimization problems. One of our main purposes is to show that these additional critical points will not introduce spurious local minima. More precisely, we want to figure out what properties of the convex function f are required for the factored objective functions g to have no spurious local minima.

4.1.2 Enlightening Examples

To gain some intuition about the properties of f such that the factored objective function g has no spurious local minima (which is one of the main goals considered in this work), let us consider the following two examples: Weighted principal component analysis (weighted PCA) and the matrix sensing problem.

4.1.2.1 Weighted PCA

Consider the symmetric weighted PCA problem in which the lifted objective function is

$$f(\mathbf{X}) = \frac{1}{2}\|\mathbf{W} \odot (\mathbf{X} - \mathbf{X}^\star)\|_F^2,$$

where $\odot$ is the Hadamard product, $\mathbf{X}^\star$ is the global optimum we want to recover and $\mathbf{W}$ is the known weighting matrix (which is assumed to have no zero entries for simplicity). After applying the Burer-Monteiro parameterization

to $f(\mathbf{X})$, we obtain the factored objective function

$$g(\mathbf{U}) = \frac{1}{2}\|\mathbf{W} \odot (\mathbf{U}\mathbf{U}^\top - \mathbf{X}^\star)\|_F^2.$$

To investigate the conditions under which the bi-linearity $\phi(\mathbf{U}) = \mathbf{U}\mathbf{U}^\top$ will (not) introduce additional local minima to the factored optimization problems, consider a simple (but enlightening) two-dimensional example where $\mathbf{W} = \begin{bmatrix} \sqrt{1+a} & 1 \\ 1 & \sqrt{1+a} \end{bmatrix}$ for some $a \geq 0, \mathbf{X}^\star = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, and $\mathbf{U} = \begin{bmatrix} x \\ y \end{bmatrix}$ for unknowns x, y. Then the factored objective function becomes

$$g(\mathbf{U}) = \frac{1+a}{2}\left(x^2-1\right)^2 + \frac{1+a}{2}\left(y^2-1\right)^2 + (xy-1)^2. \tag{4.1}$$

In this particular setting, we will see that the value of a in the weighting matrix is the deciding factor for the occurrence of spurious local minima.

Claim 4.1.1. *The factored objective function $g(\mathbf{U})$ in* (4.1) *has no spurious local minima when $a \in [0, 2)$; while for $a > 2$, spurious local minima will appear.*

Proof. First of all, we compute the gradient $\nabla g(\mathbf{U})$ and Hessian $\nabla^2 g(\mathbf{U})$:

$$\nabla g(\mathbf{U}) = 2\begin{bmatrix} (a+1)\left(x^2-1\right)x + y(xy-1) \\ (a+1)\left(y^2-1\right)y + x(xy-1) \end{bmatrix},$$
$$\nabla^2 g(\mathbf{U}) = 2\begin{bmatrix} y^2 + \left(3x^2-1\right)(a+1) & 2xy-1 \\ 2xy-1 & x^2 + \left(3y^2-1\right)(a+1) \end{bmatrix}.$$

Now we collect all the critical points by solving $\nabla g(\mathbf{U}) = 0$ and list the Hessian of g at these points as follows[5]

① $\mathbf{U}_1 = (0,0)$, $\nabla^2 g(\mathbf{U}_1) = -2\begin{bmatrix} a+1 & 1 \\ 1 & a+1 \end{bmatrix}$;

② $\mathbf{U}_2 = (1,1)$, $\nabla^2 g(\mathbf{U}_2) = 2\begin{bmatrix} 2a+3 & 1 \\ 1 & 2a+3 \end{bmatrix}$;

③ $\mathbf{U}_3 = (\sqrt{\frac{a}{a+2}}, -\sqrt{\frac{a}{a+2}})$, $\nabla^2 g(\mathbf{U}_3) = \begin{bmatrix} 4a + \frac{8}{a+2} - 6 & \frac{8}{a+2} - 6 \\ \frac{8}{a+2} - 6 & 4a + \frac{8}{a+2} - 6 \end{bmatrix}$;

④ $\mathbf{U}_4 = (\frac{\sqrt{\frac{\sqrt{a^2-4}+a}{a}}}{\sqrt{2}}, -\frac{\sqrt{2}}{a\sqrt{\frac{\sqrt{a^2-4}+a}{a}}})$, $\nabla^2 g(\mathbf{U}_4) = \begin{bmatrix} a + 3\sqrt{a^2-4} + 2 + \frac{2\sqrt{a^2-4}}{a} & -\frac{2(a+2)}{a} \\ -\frac{2(a+2)}{a} & a - 3\sqrt{a^2-4} + 2 - \frac{2\sqrt{a^2-4}}{a} \end{bmatrix}$.

[5]Note that if $\mathbf{U}$ is a critical point, so is $-\mathbf{U}$, since $\nabla g(-\mathbf{U}) = -\nabla g(\mathbf{U})$. Hence we only list one part of these critical points.

Note that the critical point $\mathbf{U}_4$ exists only for $a \geq 2$. By checking the signs of the two eigenvalues (denoted by λ_1 and λ_2) of these Hessians, we can further classify these critical points as a local minimum, a local maximum, or a saddle point[6]:

① $\lambda_1 = -2(a+2), \lambda_2 = -2a$. So, $\mathbf{U}_1$ is a local maximum for $a > 0$ and a strict saddle for $a = 0$ (see Definition 4.1.3).

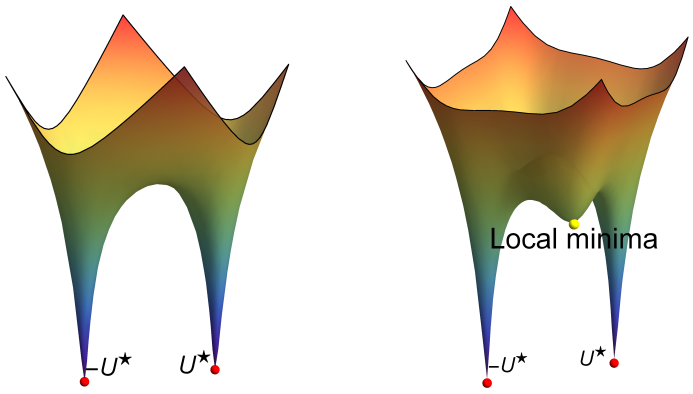
② $\lambda_1 = 4(a+1) > 0, \lambda_2 = 4(a+2) > 0$. So, $\mathbf{U}_2$ is a local minimum (also a global minimum as $g(\mathbf{U}_2) = 0$).

③ $\lambda_1 = \frac{4(a-2)(a+1)}{a+2} \begin{cases} < 0, & a \in [0, 2) \\ > 0, & a > 2 \end{cases}, \lambda_2 = 4a > 0$. So, $\mathbf{U}_3$ is $\begin{cases} \text{a saddle point}, & a \in [0, 2) \\ \text{a } \textit{spurious} \text{ local minimum}, & a > 2 \end{cases}$

④ From the determinant, we have $\lambda_1 \cdot \lambda_2 = -\frac{8(a-2)(a+1)(a+2)}{a} < 0$ for $a > 2$. So, $\mathbf{U}_4$ is a saddle point for $a > 2$.

□

In this example, the value of a controls the dynamic range of the weights as $\max W_{ij}^2 / \min W_{ij}^2 = 1 + a$. Therefore, Claim 4.1.1 can be interpreted as a relationship between the spurious local minima and the dynamic range: if the dynamic range $\max W_{ij}^2 / \min W_{ij}^2$ is smaller than 3, there will be no spurious local minima; while if the dynamic range is larger than 3, spurious local minima will appear. We also plot the landscapes of the factored objective function $g(\mathbf{U})$ in (4.1) with different dynamic ranges in Figure 4.1.



(a) Small dynamic range (b) Large dynamic range

Figure 4.1: Factored function landscapes corresponding to different dynamic ranges of the weights $\mathbf{W}$: (a) a small dynamic range with $\max W_{ij}^2 / \min W_{ij}^2 = 1$ and (b) a large dynamic range with $\max W_{ij}^2 / \min W_{ij}^2 > 3$.

As we have seen, the dynamic range of the weighting matrix serves as a determinant factor for the appearance of the spurious local minima for $g(\mathbf{U})$ in (4.1). To extend the above observations to general objective functions, we now

[6]This classification of the critical points using the Hessian information is known as the second derivative test, which says a critical point is a local maximum if the Hessian is negative definite, a local minimum is the Hessian is positive definite, and a saddle point if the Hessian matrix has both positive and negative eigenvalues.

interpret this condition (on the dynamic range of the weighting matrix) by relating it with the condition number of the Hessian matrix $\nabla^2 f(\mathbf{X})$. This can be seen from the following directional-curvature form for $f(\mathbf{X})$

$$[\nabla^2 f(\mathbf{X})](\mathbf{D}, \mathbf{D}) = \|\mathbf{W} \odot \mathbf{D}\|_F^2,$$

where $[\nabla^2 f(\mathbf{X})](\mathbf{D}, \mathbf{D})$ is the directional curvature of $f(\mathbf{X})$ along the matrix $\mathbf{D}$ of the same dimension as $\mathbf{X}$, defined by $\sum_{i,j,l,k} \frac{\partial^2 f(\mathbf{X})}{\partial X_{ij} \partial X_{lk}} D_{ij} D_{lk}$. This implies that the condition number $\lambda_{\max}(\nabla^2 f(\mathbf{X}))/\lambda_{\min}(\nabla^2 f(\mathbf{X}))$ is upper-bounded by this dynamic range:

$$\min_{ij} |W_{ij}|^2 \cdot \|\mathbf{D}\|_F^2 \le [\nabla^2 f(\mathbf{X})](\mathbf{D}, \mathbf{D}) \le \max_{ij} |W_{ij}|^2 \cdot \|\mathbf{D}\|_F^2 \quad \Longleftrightarrow \quad \frac{\lambda_{\max}(\nabla^2 f(\mathbf{X}))}{\lambda_{\min}(\nabla^2 f(\mathbf{X}))} \le \frac{\max W_{ij}^2}{\min W_{ij}^2} \tag{4.2}$$

Therefore, we conjecture that the condition number of the general convex function $f(\mathbf{X})$ would be a deciding factor of the behavior of the landscape of the factored objective function and a large condition number is very likely to introduce spurious local minima to the factored problem.

4.1.2.2 Matrix Sensing

The above conjecture can be further verified by the matrix sensing problem where the goal is to recover the low rank PSD matrix $\mathbf{X}^\star \in \mathbb{R}^{n \times n}$ from the linear measurement $\mathbf{y} = \mathcal{A}(\mathbf{X}^\star)$ with $\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ being a linear measurement operator. Consider the factored objective function $g(\mathbf{U}) = f(\mathbf{U}\mathbf{U}^\top)$ with $U \in \mathbb{R}^{n \times r}$. In [104, 125], the authors showed that the non-convex parametrization $\mathbf{U}\mathbf{U}^\top$ will not introduce spurious local minima to the factored objective function, provided the linear measurement operator $\mathcal{A}$ satisfies the following restricted isometry property (RIP).

Definition 4.1.1 (Restricted isometry property)**.** *A linear operator $\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ satisfies the r-RIP with constant δ_r if*

$$(1 - \delta_r)\|\mathbf{D}\|_F^2 \le \|\mathcal{A}(\mathbf{D})\|_2^2 \le (1 + \delta_r)\|\mathbf{D}\|_F^2 \tag{4.3}$$

holds for all $n \times n$ matrices $\mathbf{D}$ with $\operatorname{rank}(\mathbf{D}) \le r$.

Note that the required condition (4.3) essentially says that the condition number of Hessian matrix $\nabla^2 f(\mathbf{X})$ should be small at least in the directions of the low-rank matrices $\mathbf{D}$, since the directional curvature form of $f(\mathbf{X})$ is computed as $[\nabla^2 f(\mathbf{X})](\mathbf{D}, \mathbf{D}) = \|\mathcal{A}(\mathbf{D})\|_F^2$.

Inspiration. From these two examples, we see that as long as the Hessian matrix of the original convex function $f(\mathbf{X})$ has a small (restricted) condition number, the resulting factored objective function has a landscape such that all local minima correspond to the globally optimal solution. Therefore, we believe that such a restricted well-conditionedness property might be the key factor bring us a benign factored landscape, i.e.,

$$\alpha\|\mathbf{D}\|_F^2 \le [\nabla^2 f(\mathbf{X})](\mathbf{D},\mathbf{D}) \le \beta\|\mathbf{D}\|_F^2 \text{ with some small } \beta/\alpha,$$

which says that the landscape of $f(\mathbf{X})$ in the lifted space is bowl-shaped, at least in the directions of low-rank matrices.

4.1.3 Our Results

Before presenting the main results, we list a few necessary definitions.

Definition 4.1.2 (Critical points). *For a continuous function $f : \mathbb{R}^n \to \mathbb{R}$, we say $\mathbf{x} \in \mathbb{R}^n$ is a critical point of function f, if the gradient vanishes, i.e., $\nabla f(\mathbf{x}) = \mathbf{0}$.*

Definition 4.1.3 (Strict saddles; or ridable saddles [97]). *For a twice differentiable function f, a critical point $\mathbf{x}$ is a strict saddle if the Hessian matrix $\nabla^2 f(\mathbf{x})$ has at least one strictly negative eigenvalue.*

Definition 4.1.4 (Strict saddle property [100]). *A twice differentiable function satisfies strict saddle property if each critical point either corresponds to the local minima or is a strict saddle.*

Heuristically, the strict saddle property describes a geometric structure of the landscape: if a critical point is not a local minimum, then it is a strict saddle, which implies that the Hessian matrix at this point has a strictly negative eigenvalue. Hence, we can continue to decrease the function value at this point along the negative-curvature direction. This nice geometric structure ensures that many local-search algorithms, such as noisy gradient descent [114], vanilla gradient descent with random initialization [113] and the trust region method [97], can escape from all the saddle points along the directions associated with the Hessian's negative eigenvalues, and hence converge to a local minimum.

Theorem 4.1.1 (Local convergence for strict saddle property [97, 113, 114, 126, 127]). *The strict saddle property[7] allows many local-search algorithms to escape all the saddle points and converge to a local minimum.*

Our primary interest is to understand how the original convex landscapes are transformed by the factored parameterization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$ or $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, particularly how the original global optimum is mapped to the factored space, how other types of critical points are introduced, and what are their properties. To answer these questions and conclude from the previous two examples, we require that the function $f(\mathbf{X})$ in $(\mathcal{P}_0)$- $(\mathcal{P}_1)$ be restricted well-conditioned[8]:

[7] To be precise, Lee et al. [127] showed that for any function that has a Lipschitz continuous gradient and obeys the strict saddle property, first-order methods with a random initialization almost always escape all the saddle points and converge to a local minimum. The Lipschitz-gradient assumption is commonly adopted for analyzing the convergence of local-search algorithms, and we will discuss this issue after Theorem 4.3.1. To obtain explicit convergence rate, other properties (like the gradient at the points that are away from the critical points is not small) about the objective functions may be required [97, 114, 126, 128]. In this work, similar to [100], we mostly focus on the properties of the critical points, and we omit the details about the convergence rate. However, we should note that, by utilizing the similar approach in [93], it is possible to extend the strict saddle property so that we can obtain explicit convergence rate for certain algorithms [97, 114, 126] when applied for solving the factored low-rank problems.

[8] Note that the constant 1.5 for the dynamic range $\frac{\beta}{\alpha}$ in $(\mathcal{C})$ is not optimized and it is possible to slightly relax this constraint with more sophisticated analysis. However, the example of the weighted PCA in (4.1) implies that the room for improving this constant is rather limited. In particular, Claim 4.1.1 and (4.2) indicate that when $\frac{\beta}{\alpha} > 3$, the spurious local minima will occur for the weighted PCA in (4.1). Thus, as a sufficient condition for any general objective function to have no spurious local minima, a universal bound on the condition number should be at least no larger than 3, i.e., $\frac{\beta}{\alpha} \le 3$. Also aside from the lack of spurious local minima, as stated in Theorem 4.1.2, the strict saddle property is the other one that needs to be guaranteed.

$$\alpha\|\mathbf{D}\|_F^2 \le [\nabla^2 f(\mathbf{X})](\mathbf{D},\mathbf{D}) \le \beta\|\mathbf{D}\|_F^2 \text{ with } \beta/\alpha \le 1.5 \text{ whenever } \operatorname{rank}(\mathbf{X}) \le 2r \text{ and } \operatorname{rank}(\mathbf{D}) \le 4r. \quad (\mathcal{C})$$

We show that as long as the function $f(\mathbf{X})$ in the original convex programs satisfies the restricted well-conditioned assumption ($\mathcal{C}$), each critical point of the factored programs either corresponds to the low-rank globally optimal solution of the original convex programs or is a strict saddle point where the Hessian matrix $\nabla^2 g$ has a strictly negative eigenvalue. This nice geometric structure coupled with the powerful algorithmic tools provided in Theorem 4.1.1 thus allows simple iterative algorithms to solve the factored programs to a global optimum.

Theorem 4.1.2 (Informal statement of our results). *Suppose the objective function $f(\mathbf{X})$ satisfies the restricted well-conditioned assumption ($\mathcal{C}$). Assume $\mathbf{X}^\star$ is an optimal solution of $(\mathcal{P}_0)$ or $(\mathcal{P}_1)$ with $\operatorname{rank}(\mathbf{X}^\star) = r^\star$. Set $r \ge r^\star$ for the factored variables $\mathbf{U}$ and $\mathbf{V}$. Then any critical point $\mathbf{U}$ (or $(\mathbf{U},\mathbf{V})$) of the factored objective function g in $(\mathcal{F}_0)$-$(\mathcal{F}_1)$ either corresponds to the global optimum $\mathbf{X}^\star$ such that $\mathbf{X}^\star = \mathbf{U}\mathbf{U}^\top$ for $(\mathcal{P}_0)$ (or $\mathbf{X}^\star = \mathbf{U}\mathbf{V}^\top$ for $(\mathcal{P}_1)$) or is a strict saddle point (which includes a local maximum) of g.*

First note that our result covers both over-parameterization where $r > r^\star$ and exact parameterization where $r = r^\star$, while most existing results in low-rank matrix optimization problems [100, 101, 104] mainly consider the exact-parameterization case, i.e., $r = r^\star$, due to the hardness of fulfilling the gap between the metric in the factored space and the one in the lifted space for the over-parameterization case. The geometric property established in the theorem ensures that many iterative algorithms [97, 113, 114] converge to a square-root factor (or a factorization) of $\mathbf{X}^\star$, even with random initialization. Therefore, we can recover the rank-$r^\star$ global minimizer $\mathbf{X}^\star$ of $(\mathcal{P}_0)$- $(\mathcal{P}_1)$ by running local-search algorithms on the factored function $g(\mathbf{U})$ (or $g(\mathbf{U},\mathbf{V})$) if we know an upper bound on the rank $r^\star$. For problems with additional linear constraints, such as those studied in [119], one can combine the original objective function with a least-squares term that penalizes the deviation from the linear constraints. As long as the penalization parameter is large enough, the solution is equivalent to that of the constrained minimization problems and hence is also covered by our result.

4.1.4 Stylized Applications

Our main result only relies on the restricted well-conditionedness of $f(\mathbf{X})$. Therefore, in addition to low-rank matrix recovery problems [93, 100–102, 104], it is also applicable to many other low-rank matrix optimization problems with non-quadratic objective functions, including 1-bit matrix recovery, robust PCA [101], and low-rank matrix recovery with non-Gaussian noise [129]. For ease of exposition, we list the following stylized applications regarding the PSD matrices. But we note that the results listed below also hold for the cases where $\mathbf{X}$ are general nonsymmetric matrices.

4.1.4.1 Weighted PCA

We already know that in the two-dimensional case, the landscape for the factored weighted PCA problem is closely related with the dynamic range of the weighting matrix. Now we exploit Theorem 4.1.2 to derive the result for the high-dimensional case. Consider the *symmetric* weighted PCA problem where the goal is to recover the ground-truth $\mathbf{X}^\star$ from a pointwisely-weighted observation $\mathbf{Y} = \mathbf{W} \odot \mathbf{X}^\star$. Here $\mathbf{W} \in \mathbb{R}^{n\times n}$ is the known weighting matrix and the desired solution $\mathbf{X}^\star \succeq 0$ is of rank $r^\star$. A natural approach is to minimize the following squared ℓ_2 loss:

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\text{minimize}} \frac{1}{2}\|\mathbf{W} \odot (\mathbf{U}\mathbf{U}^\top - \mathbf{X}^\star)\|_F^2. \tag{4.4}$$

Unlike the low-rank approximation problem where $\mathbf{W}$ is the all-ones matrix, in general there is no analytic solutions for the weighted PCA problem (4.4) [130] and directly solving this traditional ℓ_2 loss (4.4) is known to be NP-hard [131]. We now apply Theorem 4.1.2 to the weighted PCA problem and show the objective function in (4.4) has nice geometric structures. Towards that end, define $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{W} \odot (\mathbf{X} - \mathbf{X}^\star)\|_F^2$ and compute its directional curvature as

$$[\nabla^2 f(\mathbf{X})](\mathbf{D}, \mathbf{D}) = \|\mathbf{W} \odot \mathbf{D}\|_F^2.$$

Since β/α is a restricted condition number (conditioning on directions of low-rank matrices), which must be no larger than the standard condition number $\lambda_{\max}(\nabla^2 f(\mathbf{X}))/\lambda_{\min}(\nabla^2 f(\mathbf{X}))$. Thus, together with (4.2), we have

$$\frac{\beta}{\alpha} \le \frac{\lambda_{\max}(\nabla^2 f(\mathbf{X}))}{\lambda_{\min}(\nabla^2 f(\mathbf{X}))} \le \frac{\max W_{ij}^2}{\min W_{ij}^2}.$$

Now we apply Theorem 4.1.2 to characterize the geometry of the factored problem of (4.4).

Corollary 4.1.1. *Suppose the weighting matrix* $\mathbf{W}$ *has a small dynamic range* $\frac{\max W_{ij}^2}{\min W_{ij}^2} \le 1.5$. *Then the objective function of* (4.4) *with* $r \ge r^\star$ *satisfies the strict saddle property and has no spurious local minima.*

4.1.4.2 Matrix Sensing

We now consider the matrix sensing problem which is presented before in Section 4.1.2. To apply Theorem 4.1.2, we first compare the RIP (4.3) with our restricted well-conditionedness ($\mathcal{C}$), which is copied below

$$\alpha\|\mathbf{D}\|_F^2 \le [\nabla^2 f(\mathbf{X})](\mathbf{D}, \mathbf{D}) \le \beta\|\mathbf{D}\|_F^2 \text{ with } \beta/\alpha \le 1.5 \text{ whenever } \operatorname{rank}(\mathbf{X}) \le 2r \text{ and } \operatorname{rank}(\mathbf{D}) \le 4r.$$

Clearly, the restricted well-conditionedness ($\mathcal{C}$) would hold if the linear measurement operator $\mathcal{A}$ satisfies the $4r$-RIP with a constant δ_r such that

$$\frac{1+\delta_{4r}}{1-\delta_{4r}} \le 1.5 \iff \delta_{4r} \in \left[0, \frac{1}{5}\right].$$

Now we can apply Theorem 4.1.2 to characterize the geometry of the following matrix sensing problem after the factored parameterization:

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{y}-\mathcal{A}(\mathbf{U}\mathbf{U}^\top)\|_2^2. \tag{4.5}$$

Corollary 4.1.2. *Suppose the linear map $\mathcal{A}$ satisfies the $4r$-RIP* (4.3) *with $\delta_{4r}\in[0,1/5]$. Then the objective function of* (4.5) *with $r\geq r^\star$ satisfies the strict saddle property and has no spurious local minima.*

4.1.4.3 1-bit Matrix Completion

1-bit matrix completion, as its name indicates, is the inverse problem of completing a low-rank matrix from a set of 1-bit quantized measurements

$$Y_{ij}=\mathrm{bit}(X^\star_{ij})\quad \text{for } (i,j)\in\Omega.$$

Here, $\mathbf{X}^\star\in\mathbb{R}^{n\times n}$ is the low-rank PSD matrix of rank $r^\star$, Ω is a subset of the indices $[n]\times[n]$, and $\mathrm{bit}(\cdot)$ is the 1-bit quantifier which outputs 0 or 1 in a probabilistic manner:

$$\mathrm{bit}(x)=\begin{cases}1, & \text{with probability } \sigma(x),\\ 0, & \text{with probability } 1-\sigma(x).\end{cases}$$

One typical choice for $\sigma(x)$ is the sigmoid function $\sigma(x)=\frac{e^x}{1+e^x}$. To recover $\mathbf{X}^\star$, the authors of [132] propose to minimizing the negative log-likelihood function

$$\underset{\mathbf{X}\succeq 0}{\text{minimize}}\ f(\mathbf{X}):=-\sum_{(i,j)\in\Omega}\Big[Y_{ij}\log(\sigma(X_{ij}))+(1-Y_{ij})\log(1-\sigma(X_{ij}))\Big] \tag{4.6}$$

and show that if $\|\mathbf{X}^\star\|_*\leq cn\sqrt{r^\star}$, $\max_{ij}|\mathbf{X}^\star_{ij}|\leq c$ for some small constant c, and Ω follows certain random binomial model, solving the minimization of the negative log-likelihood function with some nuclear-norm constraint would be very likely to produce a satisfying approximation to $\mathbf{X}^\star$ [132, Theorem 1].

However, when $\mathbf{X}^\star$ is extremely high-dimensional (which is the typical case in practice), it is not efficient to deal with the nuclear norm constraint and hence we propose to minimize the factored formulation of (4.6)

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\text{minimize}}\ g(\mathbf{U}):=-\sum_{(i,j)\in\Omega}\Big[Y_{ij}\log(\sigma((\mathbf{U}\mathbf{U}^\top)_{ij}))+(1-Y_{ij})\log(1-\sigma((\mathbf{U}\mathbf{U}^\top)_{ij}))\Big]. \tag{4.7}$$

In order to utilize Theorem 4.1.2 to understand the landscape of the factored objective function (4.7), we then check the following directional Hessian quadratic from of $f(\mathbf{X})$

$$[\nabla^2 f(\mathbf{X})](\mathbf{D},\mathbf{D})=\sum_{(ij)\in\Omega}\sigma'(X_{ij})D_{ij}^2.$$

For simplicity, consider the case where $\Omega = [n] \times [n]$, i.e., observe full quantized measurements. This will not increase the acquisition cost too much, since each measurement is of 1 bit. Under this assumption, we have

$$\min \sigma'(X_{ij}) \|\mathbf{D}\|_F^2 \leq [\nabla^2 f(\mathbf{X})](\mathbf{D}, \mathbf{D}) \leq \max \sigma'(X_{ij}) \|\mathbf{D}\|_F^2 \quad \iff \quad \frac{\beta}{\alpha} \leq \frac{\max \sigma'(X_{ij})}{\min \sigma'(X_{ij})}$$

Lemma 4.1.1. *Let $\Omega = [n] \times [n]$. Assume $\|\mathbf{X}\|_\infty := \max |X_{i,j}|$ is bounded by 1.3169. Then the negative log-likelihood function (4.6) $f(\mathbf{X})$ satisfies the restricted well-conditioned property.*

Proof. First of all, we claim $\sigma(x)$ is an even, positive function and decreasing when $x \geq 0$. This is because the sigmoid function $\sigma(x)$ is odd, $\sigma'(x) = \sigma(x)(1-\sigma(x)) > 0$ by $\sigma(x) \in (0,1)$, and $\sigma''(x) = -\frac{e^x(e^x-1)}{(e^x+1)^3} < 0$ for $x \geq 0$. Therefore, for any $|X_{ij}| \leq 1.3169$, we have $\frac{\max \sigma'(X_{ij})}{\min \sigma'(X_{ij})} = \frac{\max \sigma'(0)}{\min \sigma'(1.3169)} \leq 1.49995 \leq 1.5$. □

We now use Theorem 4.1.2 to characterize the landscape of the factored formulation (4.7) in the set $\mathcal{B}_{\mathbf{U}} := \{\mathbf{U} \in \mathbb{R}^{n \times r} : \|\mathbf{U}\mathbf{U}^\top\|_\infty \leq 1.3169.\}$

Corollary 4.1.3. *Set $r \geq r^\star$ in (4.7). Then the objective function (4.7) satisfies the strict saddle property and has no spurious local minima in $\mathcal{B}_{\mathbf{U}}$.*

We remark that such a constraint on $\|\mathbf{X}\|_\infty$ is also required in the seminal work [132], while by using the Burer-Monteiro parameterization, our result removes the time-consuming nuclear norm constraint.

4.1.4.4 Robust PCA

For the symmetric variant of robust PCA, the observed matrix $\mathbf{Y} = \mathbf{X}^\star + \mathbf{S}$ with $\mathbf{S}$ being sparse and $\mathbf{X}^\star$ being PSD. Traditionally, we recover $\mathbf{X}^\star$ by minimizing $||\mathbf{Y} - \mathbf{X}||_1 = \sum_{ij} |Y_{ij} - X_{ij}|$ subject to a PSD constraint. However, this formulation does not directly fit into our framework due to the non-smoothness of the ℓ_1 norm. An alternative approach is to minimize $\sum_{ij} h_a(Y_{ij} - X_{ij})$, where $h_a(.)$ is chosen to be a convex smooth approximation to the absolute value function. A possible choice is $h_a(x) = a \log((\exp(x/a) + \exp(-x/a))/2)$, which is shown to be strictly convex and smooth in [99, Lemma A.1].

4.1.4.5 Low-rank Matrix Recovery with Non-Gaussian Noise

Consider the PCA problem where the underlying noise is non-Gaussian:

$$\mathbf{Y} = \mathbf{X}^\star + \mathbf{Z},$$

i.e., the noise matrix $\mathbf{Z} \in \mathbb{R}^{n \times n}$ may not follow the Gaussian distributions. Here, $\mathbf{X}^\star \in \mathbb{R}^{n \times n}$ is a PSD matrix of rank $r^\star$. It is known that when the noise is from normal distribution, the according maximum likelihood estimator (MLE) is given by the minimizer of a squared loss function $\text{minimize}_{\mathbf{X} \succeq 0} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\|_F^2$. However, in practice, the noise is often from other distributions [133], such as Poisson, Bernoulli, Laplacian, and Cauchy, just to name a few.

In these cases, the resulting MLE, obtained by minimizing the negative log-likelihood function, is not the square loss one. Such a noise-adaptive estimator is more effective than square-loss minimization. To have a strongly convex and smooth objective function, the noise distribution should be log-strongly-concave, e.g., the Subbotin densities [129, Example 2.13], the Weibull density $f_\beta(x) = \beta x^{\beta-1}\exp(-x^\beta)$ for $\beta \geq 2$ [129, Example 2.14], and the Chernoff's density [134, Conjecture 3.1]. Once the restricted well-conditioned assumption ($\mathcal{C}$) is satisfied, we can then apply Theorem 4.1.2 to characterize the landscape of the factored formulation. Similar results apply to matrix sensing and weighted PCA when the underlying noise is non-Gaussian.

4.1.5 Prior Arts and Inspirations

Prior Arts in Non-convex Optimization Problems. The past few years have seen a surge of interest in non-convex reformulations of convex optimization problems for efficiency and scalability reasons. However, fully understanding this phenomenon, mainly the landscapes of these non-convex reformulations could be hard. Even certifying the local optimality of a point might be an NP-hard problem [135]. The existence of spurious local minima that are not global optima is a common issue [136, 137]. Also, degenerate saddle points or those surrounded by plateaus of small curvature could also prevent local-search algorithms from converging quickly to local optima [138]. Fortunately, for a range of convex optimization problems, particularly those involving low-rank matrices, the corresponding non-convex reformulations have nice geometric structures that allow local-search algorithms to converge to global optimality. Examples include low-rank matrix factorization, completion and sensing [93, 100, 101, 104], tensor decomposition and completion [114, 139], dictionary learning [99], phase retrieval [98], and many more. Based on whether smart initializations are needed, these previous works can be roughly classified into two categories. In one case, the algorithms require a problem-dependent initialization plus local refinement. A good initialization can lead to global convergence if the initial iterate lies in the attraction basin of the global optima [103, 139–141]. For low-rank matrix recovery problems, such initializations can be obtained using spectral methods [103, 140]; for other problems, it is more difficult to find an initial point located in the attraction basin [139]. The second category of works attempt to understand the empirical success of simple algorithms such as gradient descent [113], which converge to global optimality even with random initialization [93, 100, 101, 104, 113, 114]. This is achieved by analyzing the objective function's landscape and showing that they have no spurious local minima and no degenerate saddle points. Most of the works in the second category are for specific matrix sensing problems with quadratic objective functions. Our work expands this line of geometry-based convergence analysis by considering low-rank matrix optimization problems with general objective functions.

Burer-Monteiro Reformulation for PSD Matrices. In [103], the authors also considered low-rank and PSD matrix optimization problems with general objective functions. They characterized the local landscape around the global

optima, and hence their algorithms require proper initializations for global convergence. We instead characterize the global landscape by categorizing all critical points into global optima and strict saddles. This guarantees that several local-search algorithms with random initialization will converge to the global optima. Another closely related work is low-rank and PSD matrix recovery from linear observations by minimizing the factored quadratic objective function [125]. Low-rank matrix recovery from linear measurements is a particular case of our general objective function framework. Furthermore, by relating the first order optimality condition of the factored problem with the global optimality of the original convex program, our work provides a more transparent relationship between geometries of these two problems and dramatically simplifies the theoretical argument. More recently, the authors of [142] showed that for general SDPs with linear objective functions and linear constraints, the factored problems have no spurious local minimizers. In addition to showing non-existence of spurious local minimizers for general objective functions, we also quantify the curvature around the saddle points, and our result covers both over and exact parameterizations.

Burer-Monteiro Reformulation for General Matrices. The most related work is nonsymmetric matrix sensing from linear observations, which minimizes the factored quadratic objective function [106]. The ambiguity in the factored parameterization

$$\mathbf{U}\mathbf{V}^\top = (\mathbf{U}\mathbf{R})\left(\mathbf{V}\mathbf{R}^{-1^\top}\right)^\top \text{ for all nonsingular } \mathbf{R}$$

tends to make the factored quadratic objective function badly-conditioned, especially when the matrix R or its inverse is close to being singular. To overcome this problem, the regularizer

$$\Theta_E(\mathbf{U},\mathbf{V}) = \|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 \tag{4.8}$$

is proposed to ensure that $\mathbf{U}$ and $\mathbf{V}$ have almost equal energy [8, 102, 106]. In particular, with the regularizer in (4.8), it was shown in [8, 106] that $\widetilde{g}(\mathbf{U},\mathbf{V}) = f(\mathbf{U}\mathbf{V}^\top) + \mu\Theta_E(\mathbf{U},\mathbf{V})$ with a properly chosen $\mu > 0$ has similar geometric result as the one provided in Theorem 4.1.1 for $(\mathcal{P}_1)$, i.e., $\widetilde{g}(\mathbf{U},\mathbf{V})$ also obeys the strict saddle property. Compared with [8, 102, 106], our result shows that it is not necessary to introduce the extra regularization (4.8) if we solve $(\mathcal{P}_1)$ with the factorization approach. Indeed, the optimization form $\|\mathbf{X}\|_* = \min_{\mathbf{X}=\mathbf{U}\mathbf{V}^\top}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)/2$ of the nuclear norm implicitly requires $\mathbf{U}$ and $\mathbf{V}$ to have equal energy. On the other hand, we stress that our interest is to analyze the non-convex geometry of the convex problem $(\mathcal{P}_1)$ which as we explained before, has a very nice statistical performance such as it achieves minimax denoising rate [122]. Our geometrical result implies that instead of using convex solvers to solve $(\mathcal{P}_1)$, one can turn to apply local-search algorithms to solve its factored problem $(\mathcal{F}_1)$ efficiently. In this sense, as a reformulation of the convex program $(\mathcal{P}_1)$, the non-convex optimization problem $(\mathcal{F}_1)$ inherits all the statistical performance bounds for $(\mathcal{P}_1)$. Cabral et al. [143] worked on a similar problem and showed all global optima of $(\mathcal{F}_1)$ corresponds to the solution of the convex program $(\mathcal{P}_1)$. The work [144] applied the factorization approach to a more

broad class of problems. When specialized to matrix inverse problems, their results show that any local minimizer $\mathbf{U}$ and $\mathbf{V}$ with zero columns is a global minimum for the over-parameterization case, i.e., $r > \mathrm{rank}(\mathbf{X}^\star)$. However, there are no results discussing the existence of spurious local minima or the degenerate saddles in these previous works. We extend these works and further prove that as long as the loss function $f(\mathbf{X})$ is restricted well-conditioned, all local minima are global minima and there are no degenerate saddles with no requirement on the dimension of the variables. We finally note that compared with [144], our result (Theorem 4.1.2) does not depend on the existence of zero columns at the critical points and hence can provide guarantees for many local-search algorithms.

4.1.6 Notations

Denote $[n]$ as the collection of all positive integers up to n. The symbols $\mathbf{I}$ and $\mathbf{0}$ are reserved for the identity matrix and zero matrix/vector, respectively. A subscript is used to indicate its dimension when this is not clear from context. We call a matrix PSD, denoted by $\mathbf{X} \succeq 0$, if it is symmetric and all its eigenvalues are nonnegative. The notation $\mathbf{X} \succeq \mathbf{Y}$ means $\mathbf{X} - \mathbf{Y} \succeq 0$, i.e., $\mathbf{X} - \mathbf{Y}$ is PSD. The set of $r \times r$ orthogonal matrices is denoted by $\mathbb{O}_r = \{\mathbf{R} \in \mathbb{R}^{r\times r} : \mathbf{R}\mathbf{R}^\top = \mathbf{I}_r\}$. Matrix norms, such as the spectral, nuclear, and Frobenius norms, are denoted respectively by $\|\cdot\|$, $\|\cdot\|_*$ and $\|\cdot\|_F$.

The gradient of a scalar function $f(\mathbf{Z})$ with a matrix variable $\mathbf{Z} \in \mathbb{R}^{m\times n}$ is an $m \times n$ matrix, whose (i,j)th entry is $[\nabla f(\mathbf{Z})]_{i,j} = \frac{\partial f(\mathbf{Z})}{\partial Z_{ij}}$ for $i \in [m]$, $j \in [n]$. Alternatively, we can view the gradient as a linear form $[\nabla f(\mathbf{Z})](\mathbf{G}) = \langle \nabla f(\mathbf{Z}), \mathbf{G}\rangle = \sum_{i,j} \frac{\partial f(\mathbf{Z})}{\partial Z_{ij}} G_{ij}$ for any $\mathbf{G} \in \mathbb{R}^{m\times n}$. The Hessian of $f(\mathbf{Z})$ can be viewed as a 4th order tensor of dimension $m \times n \times m \times n$, whose (i,j,k,l)th entry is $[\nabla^2 f(\mathbf{Z})]_{i,j,k,l} = \frac{\partial^2 f(\mathbf{Z})}{\partial Z_{ij}\partial Z_{k,l}}$ for $i,k \in [m]$, $j,l \in [n]$. Similar to the linear form representation of the gradient, we can view the Hessian as a bilinear form defined via $[\nabla^2 f(\mathbf{Z})](\mathbf{G},\mathbf{H}) = \sum_{i,j,k,l} \frac{\partial^2 f(\mathbf{Z})}{\partial Z_{ij}\partial Z_{kl}} G_{ij}H_{kl}$ for any $\mathbf{G},\mathbf{H} \in \mathbb{R}^{m\times n}$. Yet another way to represent the Hessian is as an $mn \times mn$ matrix $[\nabla^2 f(\mathbf{Z})]_{i,j} = \frac{\partial^2 f(\mathbf{Z})}{\partial z_i \partial z_j}$ for $i,j \in [mn]$, where z_i is the ith entry of the vectorization of $\mathbf{Z}$. We will use these representations interchangeably whenever the specific form can be inferred from context. For example, in the restricted well-conditionedness assumption ($\mathcal{C}$), the Hessian is apparently viewed as an $n^2 \times n^2$ matrix and the identity $\mathbf{I}$ is of dimension $n^2 \times n^2$.

For a matrix-valued function $\phi : \mathbb{R}^{p\times q} \to \mathbb{R}^{m\times n}$, it is notationally easier to represent its gradient (or Jacobian) and Hessian as multi-linear operators. For example, the gradient, as a linear operator from $\mathbb{R}^{p\times q}$ to $\mathbb{R}^{m\times n}$, is defined via $[\nabla[\phi(\mathbf{U})](\mathbf{G})]_{ij} = \sum_{k\in[p],l\in[q]} \frac{\partial[\phi(\mathbf{U})]_{ij}}{\partial U_{kl}} G_{kl}$ for $i \in [m], j \in [n]$ and $\mathbf{G} \in \mathbb{R}^{p\times q}$; the Hessian, as a bilinear operator from $\mathbb{R}^{p\times q} \times \mathbb{R}^{p\times q}$ to $\mathbb{R}^{m\times n}$, is defined via $[\nabla^2[\phi(\mathbf{U})](\mathbf{G},\mathbf{H})]_{ij} = \sum_{k_1,k_2\in[p],l_1,l_2\in[q]} \frac{\partial^2[\phi(\mathbf{U})]_{ij}}{\partial U_{k_1l_1}\partial U_{k_2l_2}} G_{k_1l_1}H_{k_2l_2}$ for $i \in [m], j \in [n]$ and $\mathbf{G},\mathbf{H} \in \mathbb{R}^{p\times q}$. Using this notation, the Hessian of the scalar function $f(\mathbf{Z})$ of the previous paragraph, which is also the gradient of $\nabla f(\mathbf{Z}) : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$, can be viewed as a linear operator from $\mathbb{R}^{m\times m}$ to $\mathbb{R}^{m\times n}$ denoted by $[\nabla^2 f(\mathbf{Z})](\mathbf{G})$ and satisfies $\langle [\nabla^2 f(\mathbf{Z})](\mathbf{G})], \mathbf{H}\rangle = [\nabla^2 f(\mathbf{Z})](\mathbf{G},\mathbf{H})$ for $\mathbf{G},\mathbf{H} \in \mathbb{R}^{m\times n}$.

4.2 Problem Formulation

This work considers two problems: (i) the minimization of a general convex function $f(\mathbf{X})$ with the domain being positive semi-definite matrices; (ii) the minimization of a general convex function $f(\mathbf{X})$ regularized by the matrix nuclear norm $\|\mathbf{X}\|_*$ with the domain being general matrices. Let $\mathbf{X}^\star$ be an optimal solution of $(\mathcal{P}_0)$ or $(\mathcal{P}_1)$ of rank $r^\star$. To develop faster and scalable algorithms, we apply Burer-Monteiro style parameterization [119] to the low-rank optimization variable $\mathbf{X}$ in $(\mathcal{P}_0)$- $(\mathcal{P}_1)$:

$$\begin{aligned} \text{For symmetric case:} \quad & X = \phi(\mathbf{U}) := \mathbf{U}\mathbf{U}^\top \\ \text{For nonsymmetric case:} \quad & X = \psi(\mathbf{U}, \mathbf{V}) := \mathbf{U}\mathbf{V}^\top \end{aligned}$$

where $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{V} \in \mathbb{R}^{m\times r}$ with $r \geq r^\star$. With the optimization variable $\mathbf{X}$ being parameterized, the convex programs are transformed into the factored problems $(\mathcal{F}_0)$- $(\mathcal{F}_1)$:

$$\begin{aligned} \text{For symmetric case:} \quad & \underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\text{minimize}}\, g(\mathbf{U}) = f(\phi(\mathbf{U})) \\ \text{For nonsymmetric case:} \quad & \underset{\mathbf{U}\in\mathbb{R}^{n\times r}, V\in\mathbb{R}^{m\times r}}{\text{minimize}}\, g(\mathbf{U}, \mathbf{V}) = f(\psi(\mathbf{U}, \mathbf{V})) + \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right) \end{aligned}$$

Inspired by the lifting technique in constructing SDP relaxations, we refer to the variable $\mathbf{X}$ as the lifted variable, and the variables $\mathbf{U}, \mathbf{V}$ as the factored variables. Similar naming conventions apply to the optimization problems, their domains, and objective functions.

4.2.1 Consequences of the Restricted Well-conditionedness Assumption

First the restricted well-conditionedness assumption reduces to (4.3) when the objective function is quadratic. Moreover, the restricted well-conditioned assumption $(\mathcal{C})$ is similar to (4.3) in that the operator $\frac{2}{\beta+\alpha}[\nabla^2 f(\mathbf{X})]$ preserves geometric structure for low-rank matrices:

Proposition 4.2.1. *Let $f(\mathbf{X})$ satisfy the restricted well-conditionedness assumption $(\mathcal{C})$. Then*

$$\left|\frac{2}{\beta+\alpha}[\nabla^2 f(\mathbf{X})](\mathbf{G}, \mathbf{H}) - \langle \mathbf{G}, \mathbf{H}\rangle\right| \leq \frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{G}\|_F\|\mathbf{H}\|_F \leq \frac{1}{5}\|\mathbf{G}\|_F\|\mathbf{H}\|_F \tag{4.9}$$

for any matrices $\mathbf{X}, \mathbf{G}, \mathbf{H}$ of rank at most $2r$.

Proof. We extend the argument in [78] to a general function $f(\mathbf{X})$. If either $\mathbf{G}$ or $\mathbf{H}$ is zero, (4.9) holds since both sides are 0. For nonzero $\mathbf{G}$ and $\mathbf{H}$, we can assume $\|\mathbf{G}\|_F = \|\mathbf{H}\|_F = 1$ without loss of generality[9]. Then the assumption $(\mathcal{C})$ implies

[9] Otherwise, we can divide both sides of the equation (4.9) by $\|\mathbf{G}\|_F\|\mathbf{H}\|_F$ and use the homogeneity to get an equivalent version of Proposition 4.2.1 with $\mathbf{G} = \mathbf{G}/\|\mathbf{G}\|_F$ and $\mathbf{H} = \mathbf{H}/\|\mathbf{H}\|_F$, i.e., $\|\mathbf{G}\|_F = \|\mathbf{H}\|_F = 1$.

$$\alpha \|\mathbf{G}-\mathbf{H}\|_F^2 \le [\nabla^2 f(\mathbf{X})](\mathbf{G}-\mathbf{H},\mathbf{G}-\mathbf{H}) \le \beta \|\mathbf{G}-\mathbf{H}\|_F^2,$$
$$\alpha \|\mathbf{G}+\mathbf{H}\|_F^2 \le [\nabla^2 f(\mathbf{X})](\mathbf{G}+\mathbf{H},\mathbf{G}+\mathbf{H}) \le \beta \|\mathbf{G}+\mathbf{H}\|_F^2.$$

Thus we have

$$\left|2\left[\nabla^2 f(\mathbf{X})\right](\mathbf{G},\mathbf{H}) - (\beta+\alpha)\langle \mathbf{G},\mathbf{H}\rangle\right| \le \frac{\beta-\alpha}{2}\underbrace{\left(\|\mathbf{G}\|_F^2+\|\mathbf{H}\|_F^2\right)}_{=2} = \beta-\alpha = (\beta-\alpha)\underbrace{\|\mathbf{G}\|_F\|\mathbf{H}\|_F}_{=1}.$$

We complete the proof by dividing both sides by $\beta+\alpha$:

$$\left|\frac{2}{\beta+\alpha}[\nabla^2 f(\mathbf{X})](\mathbf{G},\mathbf{H}) - \langle \mathbf{G},\mathbf{H}\rangle\right| \le \frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{G}\|_F\|\mathbf{H}\|_F \le \frac{\beta/\alpha-1}{\beta/\alpha+1}\|\mathbf{G}\|_F\|\mathbf{H}\|_F \le \frac{1}{5}\|\mathbf{G}\|_F\|\mathbf{H}\|_F,$$

where in the last inequality we use the assumption that $\beta/\alpha \le 1.5$. □

Another immediate consequence of this assumption is that if the original convex program $(\mathcal{P}_0)$ has an optimal solution $\mathbf{X}^\star$ with $\mathrm{rank}(\mathbf{X}^\star) \le r$, then there is no other optimum of $(\mathcal{P}_0)$ of rank less than or equal to r:

Proposition 4.2.2. *Suppose the function $f(\mathbf{X})$ satisfies the restricted well-conditionedness ($\mathcal{C}$). Let $\mathbf{X}^\star$ be an optimum of $(\mathcal{P}_0)$ with $\mathrm{rank}(\mathbf{X}^\star) \le r$. Then $\mathbf{X}^\star$ is the unique global optimum of $(\mathcal{P}_0)$ of rank at most r.*

Proof. For the sake of a contradiction, suppose there exists another optimum $\mathbf{X}$ of $(\mathcal{P}_0)$ with $\mathrm{rank}(\mathbf{X}) \le r$ and $\mathbf{X} \neq \mathbf{X}^\star$. We begin with the second order Taylor expansion, which reads

$$f(\mathbf{X}) = f(\mathbf{X}^\star) + \langle \nabla f(\mathbf{X}^\star), \mathbf{X}-\mathbf{X}^\star\rangle + \frac{1}{2}[\nabla^2 f(t\mathbf{X}^\star+(1-t)\mathbf{X})](\mathbf{X}-\mathbf{X}^\star,\mathbf{X}-\mathbf{X}^\star),$$

for some $t \in [0,1]$. The KKT conditions for the convex optimization problem $(\mathcal{P}_0)$ states that $\nabla f(\mathbf{X}^\star) \succeq 0$ and $\nabla f(\mathbf{X}^\star)\mathbf{X}^\star = \mathbf{0}$, implying that the second term in the above Taylor expansion

$$\langle \nabla f(\mathbf{X}^\star), \mathbf{X}-\mathbf{X}^\star\rangle = \langle \nabla f(\mathbf{X}^\star), \mathbf{X}\rangle \ge 0,$$

since $\mathbf{X}$ is feasible and hence PSD. Further, since $\mathrm{rank}(t\mathbf{X}^\star+(1-t)\mathbf{X}) \le \mathrm{rank}(\mathbf{X}) + \mathrm{rank}(\mathbf{X}^\star) \le 2r$ and similarly $\mathrm{rank}(\mathbf{X}-\mathbf{X}^\star) \le 2r < 4r$, then from the restricted well-conditionedness assumption ($\mathcal{C}$) we have

$$[\nabla^2 f(\tilde{\mathbf{X}})](\mathbf{X}-\mathbf{X}^\star,\mathbf{X}-\mathbf{X}^\star) \ge \alpha\|\mathbf{X}-\mathbf{X}^\star\|_F^2.$$

Combining all, we obtain a contradiction when $\mathbf{X} \neq \mathbf{X}^\star$:

$$f(\mathbf{X}) \ge f(\mathbf{X}^\star) + \frac{1}{2}\alpha\|\mathbf{X}-\mathbf{X}^\star\|_F^2 \ge f(\mathbf{X}) + \frac{1}{2}\alpha\|\mathbf{X}-\mathbf{X}^\star\|_F^2 > f(\mathbf{X}).$$

where the second inequality follows from the optimality of $\mathbf{X}^\star$ and the third inequality holds for any $\mathbf{X} \neq \mathbf{X}^\star$. $\square$

At a high-level, the proof essentially depends on the restricted strongly convexity of the objective function of the convex program $(\mathcal{P}_0)$, which is guaranteed by the restricted well-conditionedness assumption $(\mathcal{C})$ on $f(\mathbf{X})$. The similar argument holds for $(\mathcal{P}_1)$ by noting that the sum of a (restricted) strongly convex function and a standard convex function is still (restricted) strongly convex. However, showing this requires a slightly more complicated argument due to the non-smoothness of $\|\mathbf{X}\|_*$ around those nonsingular matrices. Mainly, we need to use the concept of subgradient.

Proposition 4.2.3. *Suppose the function $f(\mathbf{X})$ satisfies the restricted well-conditionedness $(\mathcal{C})$. Let $\mathbf{X}^\star$ be a global optimum of $(\mathcal{P}_1)$ with $\mathrm{rank}(\mathbf{X}^\star) \leq r$. Then $\mathbf{X}^\star$ is the unique global optimum of $(\mathcal{P}_1)$ of rank at most r.*

Proof. For the sake of contradiction, suppose that there exists another optimum $\mathbf{X}$ of $(\mathcal{P}_1)$ with $\mathrm{rank}(\mathbf{X}) \leq r$ and $\mathbf{X} \neq \mathbf{X}^\star$. We begin with the second order Taylor expansion of $f(\mathbf{X})$, which reads

$$f(\mathbf{X}) = f(\mathbf{X}^\star) + \langle \nabla f(\mathbf{X}^\star), \mathbf{X} - \mathbf{X}^\star \rangle + \frac{1}{2}[\nabla^2 f(t\mathbf{X}^\star + (1-t)\mathbf{X})](\mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star)$$

for some $t \in [0, 1]$. From the convexity of $\|\mathbf{X}\|_*$, for any $\mathbf{D} \in \partial\|\mathbf{X}^\star\|_*$, we also have

$$\|\mathbf{X}\|_* \geq \|\mathbf{X}^\star\|_* + \langle \mathbf{D}, \mathbf{X} - \mathbf{X}^\star \rangle.$$

Combining both, we obtain

$$\begin{aligned}
f(\mathbf{X}) + \lambda\|\mathbf{X}\|_* &\overset{①}{\geq} f(\mathbf{X}^\star) + \lambda\|\mathbf{X}^\star\|_* + \langle \nabla f(\mathbf{X}^\star) + \lambda\mathbf{D}, \mathbf{X} - \mathbf{X}^\star \rangle + \frac{1}{2}[\nabla^2 f(t\mathbf{X}^\star + (1-t)\mathbf{X})](\mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star) \\
&\overset{②}{\geq} f(\mathbf{X}^\star) + \lambda\|\mathbf{X}^\star\|_* + \frac{1}{2}[\nabla^2 f(t\mathbf{X}^\star + (1-t)\mathbf{X})](\mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star) \\
&\overset{③}{\geq} f(\mathbf{X}^\star) + \lambda\|\mathbf{X}^\star\|_* + \frac{1}{2}\alpha\|\mathbf{X} - \mathbf{X}^\star\|_F^2 \\
&\overset{④}{=} f(\mathbf{X}) + \lambda\|\mathbf{X}\|_* + \frac{1}{2}\alpha\|\mathbf{X} - \mathbf{X}^\star\|_F^2 \\
&\overset{⑤}{>} f(\mathbf{X}) + \lambda\|\mathbf{X}\|_*,
\end{aligned}$$

where ① holds for any $\mathbf{D} \in \partial\|\mathbf{X}^\star\|_*$. For ②, we use fact that $\partial f_1 + \partial f_2 = \partial(f_1 + f_2)$ for any convex functions f_1, f_2, to obtain that $\nabla f(\mathbf{X}^\star) + \lambda\partial\|\mathbf{X}^\star\|_* = \partial(f(\mathbf{X}^\star) + \lambda\|\mathbf{X}^\star\|_*)$, which includes $\mathbf{0}$ since $\mathbf{X}^\star$ is a global optimum of $(\mathcal{P}_1)$. Therefore, ② follows by choosing $\mathbf{D} \in \partial\|\mathbf{X}^\star\|_*$ such that $\nabla f(\mathbf{X}^\star) + \lambda\mathbf{D} = \mathbf{0}$. ③ uses the restricted well-conditionedness assumption $(\mathcal{C})$ as $\mathrm{rank}(t\mathbf{X}^\star + (1-t)\mathbf{X}) \leq 2r$ and $\mathrm{rank}(\mathbf{X} - \mathbf{X}^\star) \leq 4r$. ④ comes from the assumption that both $\mathbf{X}$ and $\mathbf{X}^\star$ are global optimal solutions of $(\mathcal{P}_1)$. ⑤ uses the assumption that $\mathbf{X} \neq \mathbf{X}^\star$. $\square$

4.3 Understanding the Factored Landscapes for PSD Matrices

In the convex program $(\mathcal{P}_0)$, we minimize a convex function $f(\mathbf{X})$ over the PSD cone. Let $\mathbf{X}^\star$ be an optimal solution of $(\mathcal{P}_0)$ of rank $r^\star$. We re-parameterize the low-rank PSD variable $\mathbf{X}$ as

$$\mathbf{X} = \phi(\mathbf{U}) = \mathbf{U}\mathbf{U}^\top$$

where $\mathbf{U} \in \mathbb{R}^{n\times r}$ with $r \geq r^\star$ is a rectangular, matrix square root of $\mathbf{X}$. After this parametrization, the convex program is transformed into the factored problem $(\mathcal{F}_0)$ whose objective function is $g(\mathbf{U}) = f(\phi(\mathbf{U}))$.

4.3.1 Transforming the Landscape for PSD Matrices

Our primary interest is to understand how the landscape of the lifted objective function $f(\mathbf{X})$ is transformed by the factored parameterization $\phi(\mathbf{U}) = \mathbf{U}\mathbf{U}^\top$, particularly how its global optimum is mapped to the factored space, how other types of critical points are introduced, and what their properties are.

We show that if the function $f(\mathbf{X})$ is restricted well-conditioned, then each critical point of the factored objective function $g(\mathbf{U})$ in $(\mathcal{F}_0)$ either corresponds to the low-rank global solution of the original convex program $(\mathcal{P}_0)$ or is a strict saddle where the Hessian $\nabla^2 g(\mathbf{U})$ has a strictly negative eigenvalue. This implies that the factored objective function $g(\mathbf{U})$ satisfies the strict saddle property.

Theorem 4.3.1 (**Transforming the landscape for PSD matrices).** *Suppose the function $f(\mathbf{X})$ in $(\mathcal{P}_0)$ is twice continuously differentiable and is restricted well-conditioned ($\mathcal{C}$). Assume $\mathbf{X}^\star$ is an optimal solution of $(\mathcal{P}_0)$ with $\operatorname{rank}(\mathbf{X}^\star) = r^\star$. Set $r \geq r^\star$ in $(\mathcal{F}_0)$. Let $\mathbf{U}$ be any critical point of $g(\mathbf{U})$ satisfying $\nabla g(\mathbf{U}) = \mathbf{0}$. Then $\mathbf{U}$ either corresponds to a square-root factor of $\mathbf{X}^\star$, i.e.,*

$$\mathbf{X}^\star = \mathbf{U}\mathbf{U}^\top;$$

or is a strict saddle of the factored problem $(\mathcal{F}_0)$. More precisely, let $\mathbf{U}^\star \in \mathbb{R}^{n\times r}$ such that $\mathbf{X}^\star = \mathbf{U}^\star\mathbf{U}^{\star\top}$ and set $\mathbf{D} = \mathbf{U} - \mathbf{U}^\star\mathbf{R}$ with $\mathbf{R} = \arg\min_{\mathbf{R}:\mathbf{R}\in\mathbb{O}_r} \|\mathbf{U} - \mathbf{U}^\star\mathbf{R}\|_F^2$, then the curvature of $\nabla^2 g(\mathbf{U})$ along $\mathbf{D}$ is strictly negative:

$$[\nabla^2 g(\mathbf{U})](\mathbf{D},\mathbf{D}) \leq \begin{cases} -0.24\alpha \min\left\{\rho(\mathbf{U})^2, \rho(\mathbf{X}^\star)\right\} \|\mathbf{D}\|_F^2 & \text{when } r > r^\star; \\ -0.19\alpha\rho(\mathbf{X}^\star)\|\mathbf{D}\|_F^2 & \text{when } r = r^\star; \\ -0.24\alpha\rho(\mathbf{X}^\star)\|\mathbf{D}\|_F^2 & \text{when } \mathbf{U} = \mathbf{0} \end{cases}$$

with $\rho(\cdot)$ denoting the smallest nonzero singular value of its argument. This further implies

$$\lambda_{\min}(\nabla^2 g(\mathbf{U})) \leq \begin{cases} -0.24\alpha \min\left\{\rho(\mathbf{U})^2, \rho(\mathbf{X}^\star)\right\} & \textit{when } r > r^\star; \\ -0.19\alpha\rho(\mathbf{X}^\star) & \textit{when } r = r^\star; \\ -0.24\alpha\rho(\mathbf{X}^\star) & \textit{when } \mathbf{U} = \mathbf{0}. \end{cases}$$

Several remarks follow. First, the matrix $\mathbf{D}$ is the direction from the saddle point $\mathbf{U}$ to its closest globally optimal factor $\mathbf{U}^\star\mathbf{R}$ of the same dimension as $\mathbf{U}$. Second, our result covers both over-parameterization where $r > r^\star$ and exact parameterization where $r = r^\star$. Third, we can recover the rank-$r^\star$ global minimizer $\mathbf{X}^\star$ of $(\mathcal{P}_0)$ by running local-search algorithms on the factored function $g(\mathbf{U})$ if we know an upper bound on the rank $r^\star$. In particular, to apply the results in [127] where the first-order algorithms are proved to escape all the strict saddles, aside from the strict saddle property, one needs $g(\mathbf{U})$ to have a Lipschitz continuous gradient, i.e., $\|\nabla g(\mathbf{U}) - \nabla g(\mathbf{V})\|_F \leq L_c\|U - \mathbf{V}\|_F$ or $\|\nabla^2 g(\mathbf{U})\| \leq L_c$ for some positive constant L_c (also known as the Lipschitz constant). As indicated by the expression of $\nabla^2 g(\mathbf{U})$ in (4.14), it is possible that one can not find such a constant L_c for the whole space. Similar to [126] which considers the low-rank matrix factorization problem, suppose the local-search algorithm starts at $\mathbf{U}_0$ and sequentially decreases the objective value (which is true as long as the algorithm obeys certain sufficient decrease property [145]). Then it is adequate to focus on the sublevel set of g

$$\mathrm{Lev}_f(\mathbf{U}_0) = \{U : g(\mathbf{U}) \leq g(\mathbf{U}_0)\}, \tag{4.10}$$

and show that g has a Lipschitz gradient on $\mathrm{Lev}_f(\mathbf{U}_0)$. This is formally established in Proposition 4.3.1, whose proof is given in Appendix C.1.

Proposition 4.3.1. *Under the same setting as in Theorem 4.3.1, for any initial point* $\mathbf{U}_0$, $g(\mathbf{U})$ *on* $\mathrm{Lev}_f(\mathbf{U}_0)$ *defined in* (4.10) *has a Lipschitz continuous gradient with the Lipschitz constant*

$$L_c = \sqrt{2\beta\sqrt{\frac{2}{\alpha}(f(\mathbf{U}_0\mathbf{U}_0^\top) - f(\mathbf{X}^\star))} + 2\|\nabla f(\mathbf{X}^\star)\|_F + 4\beta\left(\|\mathbf{U}^\star\|_F + \frac{\sqrt{\frac{2}{\alpha}(f(\mathbf{U}_0\mathbf{U}_0^\top) - f(\mathbf{X}^\star))}}{2(\sqrt{2}-1)\rho(\mathbf{U}^\star)}\right)^2},$$

where $\rho(\cdot)$ *denotes the smallest nonzero singular value of its argument.*

4.3.2 Metrics in the Lifted and Factored Spaces

Before continuing this geometry-based argument, it is essential to have a good understanding of the domain of the factored problem and establish a metric for this domain. Since for any $\mathbf{U}$, $\phi(\mathbf{U}) = \phi(\mathbf{UR})$ where $\mathbf{R} \in \mathbb{O}_r$, the domain of the factored objective function $g(\mathbf{U})$ is stratified into equivalence classes and can be viewed as a quotient manifold [146]. The matrices in each of these equivalence classes differ by an orthogonal transformation (not

necessarily unique when the rank of $\mathbf{U}$ is less than r). One implication is that, when working in the factored space, we should consider all factorizations of $\mathbf{X}^\star$:

$$\mathcal{A}^\star = \{\mathbf{U}^\star \in \mathbb{R}^{n\times r} : \phi(\mathbf{U}^\star) = \mathbf{X}^\star\}.$$

A second implication is that when considering the distance between two points $\mathbf{U}_1$ and $\mathbf{U}_2$, one should use the distance between their corresponding equivalence classes:

$$\operatorname{dist}(\mathbf{U}_1, \mathbf{U}_2) = \min_{\mathbf{R}_1\in\mathbb{O}_r, \mathbf{R}_2\in\mathbb{O}_r} \|\mathbf{U}_1\mathbf{R}_1 - \mathbf{U}_2\mathbf{R}_2\|_F = \min_{\mathbf{R}\in\mathbb{O}_r} \|\mathbf{U}_1 - \mathbf{U}_2\mathbf{R}\|_F. \tag{4.11}$$

Under this notation, $\operatorname{dist}(\mathbf{U}, \mathbf{U}^\star) = \min_{\mathbf{R}\in\mathbb{O}_r} \|\mathbf{U} - \mathbf{U}^\star\mathbf{R}\|_F$ represents the distance between the class containing a critical point $\mathbf{U} \in \mathbb{R}^{n\times r}$ and the optimal factor class $\mathcal{A}^\star$. The second minimization problem in the definition (4.11) is known as the orthogonal Procrustes problem, where the global optimum R is characterized by the following lemma:

Lemma 4.3.1. *[147] An optimal solution for the orthogonal Procrustes problem:*

$$\mathbf{R} = \arg\min_{\tilde{\mathbf{R}}\in\mathbb{O}_r} \|\mathbf{U}_1 - \mathbf{U}_2\tilde{\mathbf{R}}\|_F^2 = \arg\max_{\tilde{\mathbf{R}}\in\mathbb{O}_r} \langle \mathbf{U}_1, \mathbf{U}_2\tilde{\mathbf{R}}\rangle$$

For any two matrices $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{n\times r}$, the following lemma relates the distance $\|\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{U}_2\mathbf{U}_2^\top\|_F$ in the lifted space to the distance $\operatorname{dist}(\mathbf{U}_1, \mathbf{U}_2)$ in the factored space. The proof is deferred to Appendix C.2.

Lemma 4.3.2. *Assume that $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{n\times r}$. Then*

$$\|\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{U}_2\mathbf{U}_2^\top\|_F \geq \min\{\rho(\mathbf{U}_1), \rho(\mathbf{U}_2)\} \operatorname{dist}(\mathbf{U}_1, \mathbf{U}_2).$$

In particular, when one matrix is of full rank, we have a similar but tighter result to relate these two distances.

Lemma 4.3.3. *[102, Lemma 5.4] Assume that $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{n\times r}$ and $\operatorname{rank}(\mathbf{U}_1) = r$. Then*

$$\|\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{U}_2\mathbf{U}_2^\top\|_F \geq 2(\sqrt{2}-1)\rho(\mathbf{U}_1) \operatorname{dist}(\mathbf{U}_1, \mathbf{U}_2).$$

4.3.3 Proof Idea: Connecting the Optimality Conditions

The proof is inspired by connecting the optimality conditions for the two programs ($\mathcal{P}_0$) and ($\mathcal{F}_0$). First of all, as the critical points of the convex optimization problem ($\mathcal{P}_0$), they are global optima and are characterized by the necessary and sufficient KKT condition [115]

$$\nabla f(\mathbf{X}^\star) \succeq 0, \nabla f(\mathbf{X}^\star)\mathbf{X}^\star = \mathbf{0}, \mathbf{X}^\star \succeq 0. \tag{4.12}$$

The factored optimization problem ($\mathcal{F}_0$) is unconstrained, with the critical points being specified by the zero gradient condition

$$\nabla g(\mathbf{U}) = 2\nabla f(\phi(\mathbf{U}))\mathbf{U} = \mathbf{0}. \tag{4.13}$$

To classify the critical points of $(\mathcal{F}_0)$, we compute the Hessian quadratic form $[\nabla^2 g(\mathbf{U})](\mathbf{D},\mathbf{D})$ as

$$[\nabla^2 g(\mathbf{U})](\mathbf{D},\mathbf{D}) = 2\langle \nabla f(\phi(\mathbf{U})), \mathbf{D}\mathbf{D}^\top\rangle + [\nabla^2 f(\phi(\mathbf{U}))](\mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top, \mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top). \tag{4.14}$$

Roughly speaking, the Hessian quadratic form has two terms – the first term involves the gradient of $f(\mathbf{X})$ and the Hessian of $\phi(\mathbf{U})$, while the second term involves the Hessian of $f(\mathbf{X})$ and the gradient of $\phi(\mathbf{U})$. Since $\phi(\mathbf{U}+\mathbf{D}) = \phi(\mathbf{U}) + \mathbf{U}\mathbf{D}^\top + \mathbf{D}\mathbf{U}^\top + \mathbf{D}\mathbf{D}^\top$, the gradient of ϕ is the linear operator $[\nabla\phi(\mathbf{U})](\mathbf{D}) = \mathbf{U}\mathbf{D}^\top + \mathbf{D}\mathbf{U}^\top$ and the Hessian bilinear operator applies as $\frac{1}{2}[\nabla^2\phi(\mathbf{U})](\mathbf{D},\mathbf{D}) = \mathbf{D}\mathbf{D}^\top$. Note in (4.14) the second quadratic form is always nonnegative since $\nabla^2 f \succeq 0$ due to the convexity of f.

For any critical point $\mathbf{U}$ of $g(\mathbf{U})$, the corresponding lifted variable $\mathbf{X} := \mathbf{U}\mathbf{U}^\top$ is PSD and satisfies $\nabla f(\mathbf{X})\mathbf{X} = \mathbf{0}$. On one hand, if $\mathbf{X}$ further satisfies $\nabla f(\mathbf{X}) \succeq 0$, then in view of the KKT conditions (4.12) and noting $\mathrm{rank}(\mathbf{X}) = \mathrm{rank}(\mathbf{U}) \leq r$, we must have $\mathbf{X} = \mathbf{X}^\star$, the global optimum of $(\mathcal{P}_0)$. On the other hand, if $\mathbf{X} \neq \mathbf{X}^\star$, implying $\nabla f(\mathbf{X}) \not\succeq 0$ due to the necessity of (4.12), then additional critical points can be introduced into the factored space. Fortunately, $\nabla f(\mathbf{X}) \not\succeq 0$ also implies that the first quadratic form in (4.14) might be negative for a properly chosen direction $\mathbf{D}$. To sum up, the critical points of $g(\mathbf{U})$ can be classified into two categories: the global optima in the optimal factor set $\mathcal{A}^\star$ with $\nabla f(\mathbf{U}\mathbf{U}^\top) \succeq 0$ and those with $\nabla f(\mathbf{U}\mathbf{U}^\top) \not\succeq 0$. For the latter case, by choosing a proper direction $\mathbf{D}$, we will argue that the Hessian quadratic form (4.14) has a strictly negative eigenvalue, and hence moving in the direction of $\mathbf{D}$ in a short distance will decrease the value of $g(\mathbf{U})$, implying that they are strict saddles and are not local minima.

We argue that a good choice of $\mathbf{D}$ is the direction from the current $\mathbf{U}$ to its closest point in the optimal factor set $\mathcal{A}^\star$. Formally, $\mathbf{D} = \mathbf{U} - \mathbf{U}^\star\mathbf{R}$ where $\mathbf{R} = \arg\min_{\mathbf{R}:R\in\mathbb{O}_r} \|\mathbf{U} - \mathbf{U}^\star\mathbf{R}\|_F$ is the optimal rotation for the orthogonal Procrustes problem. As illustrated in Figure 4.2 where we have two global solutions $\mathbf{U}^\star$ and $-\mathbf{U}^\star$ and $\mathbf{U}$ is closer to $-\mathbf{U}^\star$, the direction from $\mathbf{U}$ to $-\mathbf{U}^\star$ has more negative curvature compared to the direction from $\mathbf{U}$ to $\mathbf{U}^\star$.

Plugging this choice of $\mathbf{D}$ into the first term of (4.14), we simplify it as

$$\begin{aligned}\langle \nabla f(\mathbf{U}\mathbf{U}^\top), \mathbf{D}\mathbf{D}^\top\rangle &= \langle \nabla f(\mathbf{U}\mathbf{U}^\top), \mathbf{U}^\star\mathbf{U}^{\star\top} - \mathbf{U}^\star\mathbf{R}\mathbf{U}^\top - \mathbf{U}(\mathbf{U}^\star\mathbf{R})^\top + \mathbf{U}\mathbf{U}^\top\rangle \\ &= \langle \nabla f(\mathbf{U}\mathbf{U}^\top), \mathbf{U}^\star\mathbf{U}^{\star\top}\rangle \\ &= \langle \nabla f(\mathbf{U}\mathbf{U}^\top), \mathbf{U}^\star\mathbf{U}^{\star\top} - \mathbf{U}\mathbf{U}^\top\rangle, \end{aligned} \tag{4.15}$$

where both the second line and last line follow from the critical point property $\nabla f(\mathbf{U}\mathbf{U}^\top)\mathbf{U} = \mathbf{0}$. To gain some intuition on why (4.15) is negative while the second term in (4.14) remains small, we consider a simple example: the matrix PCA problem.

Matrix PCA Problem. Consider the PCA problem for symmetric PSD matrices

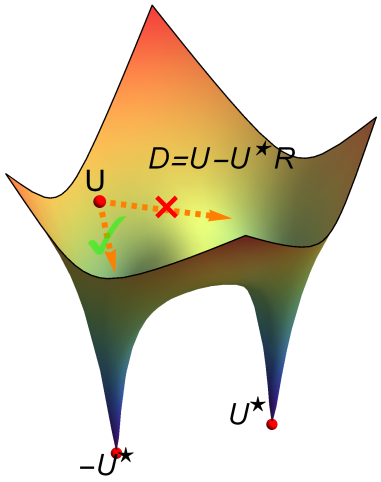


Figure 4.2: The matrix $\mathbf{D} = \mathbf{U} - \mathbf{U}^\star\mathbf{R}$ is the direction from the critical point $\mathbf{U}$ to its nearest optimal factor $\mathbf{U}^\star\mathbf{R}$, whose norm $\|\mathbf{U} - \mathbf{U}^\star\mathbf{R}\|_F$ defines the distance $\mathrm{dist}(\mathbf{U}, \mathbf{U}^\star)$. Here, $\mathbf{U}$ is closer to $-\mathbf{U}^\star$ than $\mathbf{U}^\star$ and the direction from $\mathbf{U}$ to $-\mathbf{U}^\star$ has more negative curvature compared to the direction from $\mathbf{U}$ to $\mathbf{U}^\star$.

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times n}}{\text{minimize}}\, f_{\text{PCA}}(\mathbf{X}) := \frac{1}{2}\|\mathbf{X} - \mathbf{X}^\star\|_F^2 \text{ subject to } \mathbf{X} \succeq 0, \tag{4.16}$$

where $\mathbf{X}^\star$ is a symmetric PSD matrix of rank $r^\star$. Trivially, the optimal solution is $\mathbf{X} = \mathbf{X}^\star$. Now consider the factored problem

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\text{minimize}}\, g(\mathbf{U}) := f_{\text{PCA}}(\mathbf{U}\mathbf{U}^\top) = \frac{1}{2}\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top}\|_F^2,$$

where $\mathbf{U}^\star \in \mathbb{R}^{n\times r}$ satisfies $\phi(\mathbf{U}^\star) = \mathbf{X}^\star$. Our goal is to show that any critical point $\mathbf{U}$ such that $\mathbf{X} := \mathbf{U}\mathbf{U}^\top \neq \mathbf{X}^\star$ is a strict saddle.

Controlling the first term. Since $\nabla f_{\text{PCA}}(\mathbf{X}) = \mathbf{X} - \mathbf{X}^\star$, by (4.15), the first term of $[\nabla^2 g(\mathbf{U})](\mathbf{D}, \mathbf{D})$ in (4.14) becomes

$$2\langle \nabla f_{\text{PCA}}(\mathbf{X}), \mathbf{D}\mathbf{D}^\top\rangle = 2\langle \nabla f_{\text{PCA}}(\mathbf{X}), \mathbf{X}^\star - \mathbf{X}\rangle = 2\langle \mathbf{X} - \mathbf{X}^\star, \mathbf{X}^\star - \mathbf{X}\rangle = -2\|\mathbf{X} - \mathbf{X}^\star\|_F^2, \tag{4.17}$$

which is strictly negative when $\mathbf{X} \neq \mathbf{X}^\star$.

Controlling the second term. We show that the second term $[\nabla^2 f(\phi(\mathbf{U}))](\mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top, \mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top)$ vanishes by showing that $\mathbf{D}\mathbf{U}^\top = \mathbf{0}$ (hence $\mathbf{U}\mathbf{D}^\top = \mathbf{0}$). For this purpose, let $\mathbf{X}^\star = \mathbf{Q}\,\mathrm{diag}(\boldsymbol{\lambda})\mathbf{Q}^\top = \sum_{i=1}^{r^\star} \lambda_i \mathbf{q}_i\mathbf{q}_i^\top$ be the eigenvalue decomposition of $\mathbf{X}^\star$, where $\mathbf{Q} = \begin{bmatrix}\mathbf{q}_1 & \cdots & \mathbf{q}_{r^\star}\end{bmatrix} \in \mathbb{R}^{n\times r^\star}$ has orthonormal columns and $\boldsymbol{\lambda} \in \mathbb{R}^{r^\star}$ is composed of positive entries. Similarly, let $\phi(\mathbf{U}) = \mathbf{V}\,\mathrm{diag}(\boldsymbol{\mu})\mathbf{V}^\top = \sum_{i=1}^{r'} \mu_i \mathbf{v}_i\mathbf{v}_i^\top$ be the eigenvalue decomposition of $\phi(\mathbf{U})$, where $r' = \mathrm{rank}(\mathbf{U})$. The critical point $\mathbf{U}$ satisfies $-\nabla g(\mathbf{U}) = 2(\mathbf{X}^\star - \phi(\mathbf{U}))\mathbf{U} = \mathbf{0}$, implying that

$$\mathbf{0} = \left(\mathbf{X}^\star - \sum_{i=1}^{r'} \mu_i \mathbf{v}_i \mathbf{v}_i^\top\right)\mathbf{v}_j = \mathbf{X}^\star \mathbf{v}_j - \mu_j \mathbf{v}_j, j = 1, \ldots, r'.$$

This means $(\mu_j, \mathbf{v}_j)$ forms an eigenvalue-eigenvector pair of $\mathbf{X}^\star$ for each $j = 1, \ldots, r'$. Consequently,

$$\mu_j = \lambda_{i_j} \text{ and } \mathbf{v}_j = \mathbf{q}_{i_j}, j = 1, \ldots, r'.$$

Hence $\phi(\mathbf{U}) = \sum_{j=1}^{r'} \lambda_{i_j} \mathbf{q}_{i_j} \mathbf{q}_{i_j}^\top = \sum_{j=1}^{r^\star} \lambda_j s_j \mathbf{q}_j \mathbf{q}_j^\top$. Here s_j is equal to either 0 or 1 indicating which of the eigenvalue-eigenvector pair $(\lambda_j, \mathbf{q}_j)$ appears in the decomposition of $\phi(\mathbf{U})$. Without loss of generality, we can choose $\mathbf{U}^\star = \mathbf{Q}\begin{bmatrix}\mathrm{diag}(\sqrt{\boldsymbol{\lambda}}) & \mathbf{0}\end{bmatrix}$. Then $\mathbf{U} = \mathbf{Q}\begin{bmatrix}\mathrm{diag}(\sqrt{\boldsymbol{\lambda}} \odot \mathbf{s}) & \mathbf{0}\end{bmatrix}\mathbf{V}^\top$ for some orthonormal matrix $\mathbf{V} \in \mathbb{R}^{r \times r}$ and $\mathbf{s} = \begin{bmatrix} s_1 & \cdots & s_{r^\star}\end{bmatrix}$, where the symbol $\odot$ means pointwise multiplication. By the Procrustes Lemma in [147], we obtain $\mathbf{R} = \mathbf{V}^\top$. Plugging these into $\mathbf{D}\mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star \mathbf{R} \mathbf{U}^\top$ gives $\mathbf{D}\mathbf{U}^\top = \mathbf{0}$.

Combining the two. Hence $[\nabla^2 g(\mathbf{U})](\mathbf{D}, \mathbf{D})$ is simply determined by its first term

$$\begin{aligned}
[\nabla^2 g(\mathbf{U})](\mathbf{D}, \mathbf{D}) &= -2\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star \mathbf{U}^{\star\top}\|_F^2 \\
&\leq -2 \min\left\{\rho(\mathbf{U})^2, \rho(\mathbf{U}^\star)^2\right\} \|\mathbf{D}\|_F^2 \\
&= -2 \min\left\{\rho(\phi(\mathbf{U})), \rho(\mathbf{X}^\star)\right\} \|\mathbf{D}\|_F^2 \\
&= -2\rho(\mathbf{X}^\star)\|\mathbf{D}\|_F^2,
\end{aligned}$$

where the second line follows from Lemma 4.3.2 and the last line follows from the fact that all the eigenvalues of $\mathbf{U}\mathbf{U}^\top$ come from those of $\mathbf{X}^\star$. Finally, we obtain the desired strict saddle property of $g(\mathbf{U})$:

$$\lambda_{\min}(\nabla^2 g(\mathbf{U})) \leq -2\rho(\mathbf{X}^\star).$$

This simple example is ideal in several ways, particularly the gradient $\nabla f(\phi(\mathbf{U})) = \phi(\mathbf{U}) - \phi(\mathbf{U}^\star)$, which directly establishes the negativity of the first term in (4.14); and by choosing $\mathbf{D} = \mathbf{U} - \mathbf{U}^\star \mathbf{R}$ and using $\mathbf{D}\mathbf{U}^\top = \mathbf{0}$, the second term vanishes. Neither of these simplifications hold for general objective functions $f(\mathbf{X})$. However, the example does suggest that the direction $\mathbf{D} = \mathbf{U} - \mathbf{U}^\star \mathbf{R}$ is a good choice to show $[\nabla^2 g(\mathbf{U})](\mathbf{D}, \mathbf{D}) \leq -\tau\|\mathbf{D}\|_F^2$ for some $\tau > 0$. For a formal proof, we will also use the direction $\mathbf{D} = \mathbf{U} - \mathbf{U}^\star \mathbf{R}$ to show that those critical points $\mathbf{U}$ not corresponding to $\mathbf{X}^\star$ have a negative directional curvature for the general factored objective function $g(\mathbf{U})$.

4.3.4 A Formal Proof of Theorem 4.3.1

Proof Outline. We present a formal proof of Theorem 4.3.1 in this section. The main argument involves showing each critical point $\mathbf{U}$ of $g(\mathbf{U})$ either corresponds to the optimal solution $\mathbf{X}^\star$ or its Hessian matrix $\nabla^2 g(\mathbf{U})$ has at least one strictly negative eigenvalue. Inspired by the discussions in Section 4.3.3, we will use the direction $\mathbf{D} =$

$\mathbf{U} - \mathbf{U}^\star\mathbf{R}$ and show that the Hessian $\nabla^2 g(\mathbf{U})$ has a strictly negative directional curvature in the direction of $\mathbf{D}$, i.e., $[\nabla^2 g(\mathbf{U})](\mathbf{D}, \mathbf{D}) \leq -\tau\|\mathbf{D}\|_F^2, \text{ for some } \tau > 0.$

Supporting Lemmas. We first list two lemmas. The first lemma separates $\|(\mathbf{U}-\mathbf{Z})\mathbf{U}^\top\|_F^2$ into two terms: $\|\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top\|_F^2$ and $\|(\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top)\mathbf{Q}\mathbf{Q}^\top\|_F^2$ with $\mathbf{Q}\mathbf{Q}^\top$ being the projection matrix onto $\mathrm{Range}(\mathbf{U})$. It is crucial for the first term $\|\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top\|_F^2$ to have a small coefficient. In the second lemma, we will further control the second term as a consequence of $\mathbf{U}$ being a critical point. The proof of Lemma 4.3.4 is given in Section C.3.

Lemma 4.3.4. *Let $\mathbf{U}$ and $\mathbf{Z}$ be any two matrices in $\mathbb{R}^{n\times r}$ such that $\mathbf{U}^\top\mathbf{Z} = \mathbf{Z}^\top\mathbf{U}$ is PSD. Assume that $\mathbf{Q}$ is an orthogonal matrix whose columns span $\mathrm{Range}(\mathbf{U})$. Then*

$$\left\|(\mathbf{U}-\mathbf{Z})\mathbf{U}^\top\right\|_F^2 \leq \frac{1}{8}\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top\right\|_F^2 + \left(3 + \frac{1}{2\sqrt{2}-2}\right)\left\|(\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top)\mathbf{Q}\mathbf{Q}^\top\right\|_F^2.$$

We remark that Lemma 4.3.4 is a strengthened version of [125, Lemma 4.4]. While the result there requires: (i) $\mathbf{U}$ to be a critical point of the factored objective function $g(\mathbf{U})$; (ii) $\mathbf{Z}$ to be an optimal factor in $\mathcal{A}^\star$ that is closest to $\mathbf{U}$, i.e., $\mathbf{Z} = \mathbf{U}^\star\mathbf{R}$ with $\mathbf{U}^\star \in \mathcal{A}^\star$ and $\mathbf{R} = \arg\min_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{I}_r} \|\mathbf{W} - \mathbf{W}^\star\mathbf{R}\|_F$. Lemma 4.3.4 removes these assumptions and requires only $\mathbf{U}^\top\mathbf{Z} = \mathbf{Z}^\top\mathbf{U}$ being PSD.

Next, we control the distance between $\mathbf{U}\mathbf{U}^\top$ and the global solution $\mathbf{X}^\star$ when $\mathbf{U}$ is a critical point of the factored objective function $g(\mathbf{U})$, i.e., $\nabla g(\mathbf{U}) = \mathbf{0}$. The proof, given in Section C.4, relies on writing $\nabla f(\mathbf{X}) = \nabla f(\mathbf{X}^\star) + \int_0^1 [\nabla^2 f(t\mathbf{X} + (1-t)\mathbf{X}^\star)](\mathbf{X} - \mathbf{X}^\star)dt$ and applying Proposition 4.2.1.

Lemma 4.3.5 (Upper Bound on $\|(\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\|_F^2$). *Suppose the objective function $f(\mathbf{X})$ in $(\mathcal{P}_0)$ is twice continuously differentiable and satisfies the restricted well-conditionedness assumption $(\mathcal{C})$. Further, let $\mathbf{U}$ be any critical point of $(\mathcal{F}_0)$ and $\mathbf{Q}$ be the orthonormal basis spanning $\mathrm{Range}(\mathbf{U})$. Then*

$$\left\|(\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\right\|_F \leq \frac{\beta-\alpha}{\beta+\alpha}\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top}\right\|_F.$$

Proof of Theorem 4.3.1. Along the same lines as in the matrix PCA example, it suffices to find a direction $\mathbf{D}$ to produce a strictly negative curvature for each critical point $\mathbf{U}$ not corresponding to $\mathbf{X}^\star$. We choose $\mathbf{D} = \mathbf{U} - \mathbf{U}^\star\mathbf{R}$ where $\mathbf{R} = \arg\min_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{I}_r} \|\mathbf{W} - \mathbf{W}^\star\mathbf{R}\|_F$. Then

$$
\begin{aligned}
&[\nabla^2 g(\mathbf{U})](\mathbf{D},\mathbf{D}) \\
=&2\langle \nabla f(\mathbf{X}), \mathbf{D}\mathbf{D}^\top\rangle + [\nabla^2 f(\mathbf{X})](\mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top, \mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top) && \text{By Eq. (4.14)} \\
=&2\langle \nabla f(\mathbf{X}), \mathbf{X}^\star - \mathbf{X}\rangle + [\nabla^2 f(\mathbf{X})](\mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top, \mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top) && \text{By Eq. (4.13)} \\
\leq& \underbrace{2\langle \nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^\star), \mathbf{X}^\star - \mathbf{X}\rangle}_{\Pi_1} + \underbrace{[\nabla^2 f(\mathbf{X})](\mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top, \mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top)}_{\Pi_2} && \text{By Eq. (4.12)}
\end{aligned}
$$

In the following, we will bound Π_1 and Π_2, respectively.

Bounding Π_1.

$$
\begin{aligned}
\Pi_1 = -2\langle \nabla f(\mathbf{X}^\star) - \nabla f(\mathbf{X}), \mathbf{X}^\star - \mathbf{X}\rangle &\overset{①}{=} -2\left\langle \int_0^1 [\nabla^2 f(t\mathbf{X} + (1-t)\mathbf{X}^\star)](\mathbf{X}^\star - \mathbf{X})dt, \mathbf{X}^\star - \mathbf{X}\right\rangle \\
&= -2\int_0^1 \left[\nabla^2 f(t\mathbf{X} + (1-t)\mathbf{X}^\star)\right](\mathbf{X}^\star - \mathbf{X}, \mathbf{X}^\star - \mathbf{X})dt \\
&\overset{②}{\leq} -2\alpha\|\mathbf{X}^\star - \mathbf{X}\|_F^2,
\end{aligned}
$$

where ① follows from the Taylor's Theorem for vector-valued functions [148, Eq. (2.5) in Theorem 2.1], and ② follows from the restricted strong convexity assumption ($\mathcal{C}$) since the PSD matrix $t\mathbf{X} + (1-t)\mathbf{X}^\star$ has rank of at most $2r$ and $\mathrm{rank}(\mathbf{X}^\star - \mathbf{X}) \leq 4r$.

Bounding Π_2.

$$
\begin{aligned}
\Pi_2 &= [\nabla^2 f(\mathbf{X})](\mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top, \mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top) \\
&\leq \beta\|\mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top\|_F^2 && \text{By } (\mathcal{C}) \\
&\leq 4\beta\|\mathbf{D}\mathbf{U}^\top\|_F^2 \\
&\leq 4\beta\left[\frac{1}{8}\|\mathbf{X} - \mathbf{X}^\star\|_F^2 + \left(3 + \frac{1}{2\sqrt{2}-2}\right)\|(\mathbf{X} - \mathbf{X}^\star)\mathbf{Q}\mathbf{Q}^\top\|_F^2\right]. && \text{By Lemma 4.3.4} \\
&\leq 4\beta\left[\frac{1}{8} + \left(3 + \frac{1}{2\sqrt{2}-2}\right)\frac{(\beta-\alpha)^2}{(\beta+\alpha)^2}\right]\|\mathbf{X} - \mathbf{X}^\star\|_F^2 && \text{By Lemma 4.3.5} \\
&\leq 1.76\alpha\|\mathbf{X}^\star - \mathbf{X}\|_F^2. && \text{By } \beta/\alpha \leq 1.5
\end{aligned}
$$

Combining the two. Hence,

$$
\Pi_1 + \Pi_2 \leq -0.24\alpha\|\mathbf{X}^\star - \mathbf{X}\|_F^2.
$$

Then, we relate the lifted distance $\|\mathbf{X}^\star - \mathbf{X}\|_F^2$ with the factored distance $\|\mathbf{U} - \mathbf{U}^\star\mathbf{R}\|_F^2$ using Lemma 4.3.2 when $r > r^\star$, and Lemma 4.3.3 when $r = r^\star$, respectively:

$$\begin{aligned}\text{When } r > r^\star\text{: } [\nabla^2 g(\mathbf{U})](\mathbf{D},\mathbf{D}) &\le -0.24\alpha \min\left\{\rho(\mathbf{U})^2, \rho(\mathbf{U}^\star)^2\right\}\|\mathbf{D}\|_F^2 && \text{By Lemma 4.3.2}\\ &= -0.24\alpha \min\left\{\rho(\mathbf{U})^2, \rho(\mathbf{X}^\star)\right\}\|\mathbf{D}\|_F^2;\end{aligned}$$

$$\begin{aligned}\text{When } r = r^\star\text{: } [\nabla^2 g(\mathbf{U})](\mathbf{D},\mathbf{D}) &\le -0.19\alpha\rho(\mathbf{U}^\star)^2\|\mathbf{D}\|_F^2 && \text{By Lemma 4.3.3}\\ &= -0.19\alpha\rho(\mathbf{X}^\star)\|\mathbf{D}\|_F^2.\end{aligned}$$

For the special case where $\mathbf{U} = \mathbf{0}$, we have

$$\begin{aligned}[\nabla^2 g(\mathbf{U})](\mathbf{D},\mathbf{D}) &\le -0.24\alpha\|\mathbf{0} - \mathbf{X}^\star\|_F^2\\ &= -0.24\alpha\|\mathbf{U}^\star\mathbf{U}^{\star\top}\|_F^2\\ &\le -0.24\alpha\rho(\mathbf{U}^\star)^2\|\mathbf{U}^\star\|_F^2\\ &= -0.24\alpha\rho(\mathbf{X}^\star)\|\mathbf{D}\|_F^2,\end{aligned}$$

where the last second line follows from

$$\|\mathbf{U}^\star\mathbf{U}^{\star\top}\|_F^2 = \sum_i \sigma_i^4(\mathbf{U}^\star) = \sum_{i:\sigma_i(\mathbf{U}^\star)\neq 0} \sigma_i^4(\mathbf{U}^\star) \ge \min_{i:\sigma_i(\mathbf{U}^\star)\neq 0} \sigma_i^2(\mathbf{U}^\star)\left(\sum_{j:\sigma_j(\mathbf{U}^\star)\neq 0} \sigma_j^2(\mathbf{U}^\star)\right) = \rho^2(\mathbf{U}^\star)\|\mathbf{U}^\star\|_F^2,$$

and the last line follows from $\mathbf{D} = \mathbf{0} - \mathbf{U}^\star\mathbf{R} = -\mathbf{U}^\star\mathbf{R}$ when $\mathbf{U} = \mathbf{0}$. Here $\sigma_i(\cdot)$ denotes the i-th largest singular value of its argument. □

4.4 Understanding the Factored Landscapes for General Non-square Matrices

In this section, we will study the second convex program ($\mathcal{P}_1$): the minimization of a general convex function $f(\mathbf{X})$ regularized by the matrix nuclear norm $\|\mathbf{X}\|_*$ with the domain being general matrices. Since the matrix nuclear norm $\|\mathbf{X}\|_*$ appears in the objective function, the standard convex solvers or even faster tailored ones require performing singular value decomposition in each iteration, which severely limits the efficiency and scalability of the convex program. Motivated by this, we will instead solve its Burer-Monteiro re-parameterized counterpart.

4.4.1 Burer-Monteiro Reformulation of the Nuclear Norm Regularization

Recall the second problem is the nuclear norm regularization ($\mathcal{P}_1$):

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times m}}{\text{minimize}}\ f(\mathbf{X}) + \lambda\|\mathbf{X}\|_* \tag{$\mathcal{P}_1$}$$

This convex program has an equivalent SDP formulation [65, page 8]:

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times m},\mathbf{\Phi}\in\mathbb{R}^{n\times n},\mathbf{\Psi}\in\mathbb{R}^{m\times m}}{\text{minimize}}\ f(\mathbf{X}) + \frac{\lambda}{2}(\mathrm{tr}(\mathbf{\Phi}) + \mathrm{tr}(\mathbf{\Psi}))\ \text{ subject to }\ \begin{bmatrix}\mathbf{\Phi} & \mathbf{X}\\ \mathbf{X}^\top & \mathbf{\Psi}\end{bmatrix} \succeq 0. \tag{4.18}$$

When the PSD constraint is implicitly enforced as the following equality constraint

$$\begin{bmatrix} \mathbf{\Phi} & \mathbf{X} \\ \mathbf{X}^\top & \mathbf{\Psi} \end{bmatrix} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}^\top \Rightarrow \mathbf{X} = \mathbf{U}\mathbf{V}^\top, \mathbf{\Phi} = \mathbf{U}\mathbf{U}^\top, \mathbf{\Psi} = \mathbf{V}\mathbf{V}^\top, \tag{4.19}$$

we obtain the Burer-Monteiro factored reformulation ($\mathcal{F}_1$):

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}, \mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \; g(\mathbf{U},\mathbf{V}) = f(\mathbf{U}\mathbf{V}^\top) + \frac{\lambda}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2). \tag{$\mathcal{F}_1$}$$

The factored formulation ($\mathcal{F}_1$) can potentially solve the computational issue of ($\mathcal{P}_1$) in two major respects: *(i)* avoiding expensive SVDs by replacing the nuclear norm $\|\mathbf{X}\|_*$ with the squared term $(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)/2$; *(ii)* a substantial reduction in the number of the optimization variables from nm to $(n+m)r$.

4.4.2 Transforming the Landscape for General Non-square Matrices

Our primary interest is to understand how the landscape of the lifted objective function $f(\mathbf{X}) + \lambda\|\mathbf{X}\|_*$ is transformed by the factored parameterization $\psi(\mathbf{U},\mathbf{V}) = \mathbf{U}\mathbf{V}^\top$. The main contribution of this part is establishing that under the restricted well-conditionedness of the convex loss function $f(\mathbf{X})$, the factored formulation ($\mathcal{F}_1$) has no spurious local minima and satisfies the strict saddle property.

Theorem 4.4.1 (**Transforming the landscape for general non-square matrices**). *Suppose the function $f(\mathbf{X})$ satisfies the restricted well-conditioned property ($\mathcal{C}$). Assume that $\mathbf{X}^\star$ of rank $r^\star$ is an optimal solution of ($\mathcal{P}_1$) where $\lambda > 0$. Set $r \geq r^\star$ in the factored program ($\mathcal{F}_1$). Let $(\mathbf{U},\mathbf{V})$ be any critical point of $g(\mathbf{U},\mathbf{V})$ satisfying $\nabla g(\mathbf{U},\mathbf{V}) = \mathbf{0}$. Then $(\mathbf{U},\mathbf{V})$ either corresponds to a factorization of $\mathbf{X}^\star$, i.e.,*

$$\mathbf{X}^\star = \mathbf{U}\mathbf{V}^\top;$$

or is a strict saddle of the factored problem:

$$\lambda_{\min}(\nabla^2 g(\mathbf{U},\mathbf{V})) \leq \begin{cases} -0.12\alpha \min\left\{0.5\rho^2(\mathbf{W}), \rho(\mathbf{X}^\star)\right\} & \text{when } r > r^\star; \\ -0.099\alpha\rho(\mathbf{X}^\star) & \text{when } r = r^\star; \\ -0.12\alpha\rho(\mathbf{X}^\star) & \text{when } \mathbf{W} = \mathbf{0}, \end{cases}$$

where $W := \begin{bmatrix} \mathbf{U}^\top & \mathbf{V}^\top \end{bmatrix}^\top$ and $\rho(\mathbf{W})$ is the smallest nonzero singular value of $\mathbf{W}$.

Theorem 4.4.1 ensures that many local-search algorithms[10] when applied for solving the factored program ($\mathcal{F}_1$), can escape from all the saddle points and converge to a global solution that corresponds to $\mathbf{X}^\star$. Several remarks follow.

[10]The Lipschitz gradient of g at any its sublevel set can be obtained with similar approach for Proposition 4.3.1.

The Non-triviality of Extending the PSD Case to the Nonsymmetric Case. Although the generalization from the PSD case might not seem technically challenging at first sight, we must overcome several technical difficulties to prove this main theorem. We make a few other technical contributions in the process. In fact, the non-triviality of extending to the nonsymmetric case is also highlighted in [102, 104, 106]. The major technique difficulty to complete such an extension is the ambiguity issue existed in the nonsymmetric case: $\mathbf{U}\mathbf{V}^\top = (t\mathbf{U})(1/t\mathbf{V})^\top$ for any nonzero t. This tends to make the factored quadratic objective function badly-conditioned, especially when t is very large or small. To prevent this from happening, a popular strategy utilized to adapt the result for the symmetric case to the non-symmetric case is to introduce an additional balancing regularization to ensure that $\mathbf{U}$ and $\mathbf{V}$ have equal energy [102, 104, 106]. Sometimes these additional regularizations are quite complicated (see Eq. (13)-(15) in [140]). Instead, we find for nuclear norm regularized problems, the critical points are automatically balanced even without these additional complex balancing regularizations (see Section 4.4.4 for details). In addition, by connecting the optimality conditions of the convex program $(\mathcal{P}_1)$ and the factored program $(\mathcal{F}_1)$, we dramatically simplify the proof argument, making the relationship between the original convex problem and the factored program more transparent.

Proof Sketch of Theorem 4.4.1. We try to understand how the parameterization $\mathbf{X} = \psi(\mathbf{U}, \mathbf{V})$ transforms the geometric structures of the convex objective function $f(\mathbf{X})$ by categorizing the critical points of the non-convex factored function $g(\mathbf{U}, \mathbf{V})$. In particular, we will illustrate how the globally optimal solution of the convex program is transformed in the domain of $g(\mathbf{U}, \mathbf{V})$. Furthermore, we will explore the properties of the additional critical points introduced by the parameterization and find a way of utilizing these properties to prove the strict saddle property. For those purposes, the optimality conditions for the two programs $(\mathcal{P}_1)$ and $(\mathcal{F}_1)$ will be compared.

4.4.3 Optimality Condition for the Convex Program

As an unconstrained convex optimization, all critical points of $(\mathcal{P}_1)$ are global optima and are characterized by the necessary and sufficient KKT condition [115]:

$$\nabla f(\mathbf{X}^\star) \in -\lambda \partial \|\mathbf{X}^\star\|_*, \tag{4.20}$$

where $\partial\|\mathbf{X}^\star\|_*$ denotes the subdifferential (the set of subgradient) of the nuclear norm $\|\mathbf{X}\|_*$ evaluated at $\mathbf{X}^\star$. The subdifferential of the matrix nuclear norm is defined by

$$\partial\|\mathbf{X}\|_* = \{\mathbf{D} \in \mathbb{R}^{n\times m} : \|\mathbf{Y}\|_* \geq \|\mathbf{X}\|_* + \langle \mathbf{Y} - \mathbf{X}, \mathbf{D}\rangle, \text{all } \mathbf{Y} \in \mathbb{R}^{n\times m}\}.$$

We have a more explicit characterization of the subdifferential of the nuclear norm using the singular value decomposition. More specifically, suppose $\mathbf{X} = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top$ is the (compact) singular value decomposition of $\mathbf{X} \in \mathbb{R}^{n\times m}$ with $\mathbf{P} \in \mathbb{R}^{n\times r}, \mathbf{Q} \in \mathbb{R}^{m\times r}$ and $\mathbf{\Sigma}$ being an $r \times r$ diagonal matrix. Then the subdifferential of the matrix nuclear norm at

$\mathbf{X}$ is given by [65, Equation (2.9)]

$$\partial\|\mathbf{X}\|_* = \{\mathbf{P}\mathbf{Q}^\top + \mathbf{E} : \mathbf{P}^\top\mathbf{E} = \mathbf{0}, \mathbf{E}\mathbf{Q} = \mathbf{0}, \|\mathbf{E}\| \leq 1\}.$$

Combining this representation of the subdifferential and the KKT condition (4.20) yields an equivalent expression for the optimality condition

$$\begin{aligned} &\nabla f(\mathbf{X}^\star)\mathbf{Q}^\star = -\lambda\mathbf{P}^\star, \\ &\nabla f(\mathbf{X}^\star)^\top\mathbf{P}^\star = -\lambda\mathbf{Q}^\star, \\ &\|\nabla f(\mathbf{X}^\star)\| \leq \lambda, \end{aligned} \tag{4.21}$$

where we assume the compact SVD of $\mathbf{X}^\star$ is given by

$$\mathbf{X}^\star = \mathbf{P}^\star\mathbf{\Sigma}^\star\mathbf{Q}^{\star\top} \text{ with } \mathbf{P}^\star \in \mathbb{R}^{n\times r^\star}, \mathbf{Q}^\star \in \mathbb{R}^{m\times r^\star}, \mathbf{\Sigma}^\star \in \mathbb{R}^{r^\star\times r^\star}.$$

Since $r \geq r^\star$ in the factored problem ($\mathcal{F}_1$), to match the dimensions, we define the optimal factors $\mathbf{U}^\star \in \mathbb{R}^{n\times r}$, $\mathbf{V}^\star \in \mathbb{R}^{m\times r}$ for any $\mathbf{R} \in \mathbb{O}_r$ as

$$\begin{aligned} \mathbf{U}^\star &= \mathbf{P}^\star[\sqrt{\mathbf{\Sigma}^\star}\ \mathbf{0}_{r^\star\times(r-r^\star)}]\mathbf{R}, \\ \mathbf{V}^\star &= \mathbf{Q}^\star[\sqrt{\mathbf{\Sigma}^\star}\ \mathbf{0}_{r^\star\times(r-r^\star)}]\mathbf{R}. \end{aligned} \tag{4.22}$$

Consequently, with the optimal factors $\mathbf{U}^\star, \mathbf{V}^\star$ defined in (4.22), we can rewrite the optimal condition (4.21) as

$$\begin{aligned} &\nabla f(\mathbf{X}^\star)\mathbf{V}^\star = -\lambda\mathbf{U}^\star, \\ &\nabla f(\mathbf{X}^\star)^\top\mathbf{U}^\star = -\lambda\mathbf{V}^\star, \\ &\|\nabla f(\mathbf{X}^\star)\| \leq \lambda. \end{aligned} \tag{4.23}$$

Stacking $\mathbf{U}^\star, \mathbf{V}^\star$ as $\mathbf{W}^\star = \begin{bmatrix} \mathbf{U}^\star \\ \mathbf{V}^\star \end{bmatrix}$ and defining

$$\Xi(\mathbf{X}) := \begin{bmatrix} \lambda\mathbf{I} & \nabla f(\mathbf{X}) \\ \nabla f(\mathbf{X})^\top & \lambda\mathbf{I} \end{bmatrix} \text{ for all } \mathbf{X} \tag{4.24}$$

yield a more concise form of the optimality condition:

$$\begin{aligned} &\Xi(\mathbf{X}^\star)\mathbf{W}^\star = \mathbf{0}, \\ &\|\nabla f(\mathbf{X}^\star)\| \leq \lambda. \end{aligned} \tag{4.25}$$

4.4.4 Characterizing the Critical Points of the Factored Program

To begin with, the gradient of $g(\mathbf{U}, \mathbf{V})$ can be computed and rearranged as

$$\begin{aligned}\nabla g(\mathbf{U}, \mathbf{V}) &= \begin{bmatrix}\nabla_{\mathbf{U}} g(\mathbf{U}, \mathbf{V}) \\ \nabla_{\mathbf{V}} g(\mathbf{U}, \mathbf{V})\end{bmatrix} \\ &= \begin{bmatrix}\nabla f(\mathbf{U}\mathbf{V}^\top)\mathbf{V} + \lambda \mathbf{U} \\ \nabla f(\mathbf{U}\mathbf{V}^\top)^\top \mathbf{U} + \lambda \mathbf{V}\end{bmatrix} \\ &= \begin{bmatrix}\lambda \mathbf{I} & \nabla f(\mathbf{U}\mathbf{V}^\top) \\ \nabla f(\mathbf{U}\mathbf{V}^\top)^\top & \lambda \mathbf{I}\end{bmatrix} \begin{bmatrix}\mathbf{U} \\ \mathbf{V}\end{bmatrix} \\ &= \Xi(\mathbf{U}\mathbf{V}^\top) \begin{bmatrix}\mathbf{U} \\ \mathbf{V}\end{bmatrix},\end{aligned} \tag{4.26}$$

where the last equality follows from the definition (4.24) of $\Xi(\cdot)$. Therefore, all critical points of $g(\mathbf{U}, \mathbf{V})$ can be characterized by the following set

$$\mathcal{X} := \left\{ (\mathbf{U}, \mathbf{V}) : \Xi(\mathbf{U}\mathbf{V}^\top) \begin{bmatrix}\mathbf{U} \\ \mathbf{V}\end{bmatrix} = \mathbf{0} \right\}.$$

We will see that any critical point $(\mathbf{U}, \mathbf{V}) \in \mathcal{X}$ forms an balanced pair, which is defined as follows:

Definition 4.4.1 (Balanced pairs). *We call $(\mathbf{U}, \mathbf{V})$ is a balanced pair if the Gram matrices of $\mathbf{U}$ and $\mathbf{V}$ are the same: $\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V} = \mathbf{0}$. All the balanced pairs form the balanced set, denoted by $\mathcal{E} := \left\{ (\mathbf{U}, \mathbf{V}) : \mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V} = \mathbf{0} \right\}$.*

By Definition 4.4.1, to show that each critical point forms an balanced pair, we rely on the following fact:

$$\mathbf{W} = \begin{bmatrix}\mathbf{U} \\ \mathbf{V}\end{bmatrix}, \widehat{\mathbf{W}} = \begin{bmatrix}\mathbf{U} \\ -\mathbf{V}\end{bmatrix} \text{ with } (\mathbf{U}, \mathbf{V}) \in \mathcal{E} \Leftrightarrow \widehat{\mathbf{W}}^\top \mathbf{W} = W^\top \widehat{\mathbf{W}} = \mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V} = \mathbf{0}. \tag{4.27}$$

Now we are ready to relate the critical points and balanced pairs, the proof of which is given in Appendix C.5.

Proposition 4.4.1. *Any critical point $(\mathbf{U}, \mathbf{V}) \in \mathcal{X}$ forms a balanced pair in $\mathcal{E}$.*

4.4.4.1 The Properties of the Balanced Set

In this part, we introduce some important properties of the balanced set $\mathcal{E}$. These properties basically compare the on-diagonal-block energy and the off-diagonal-block energy for a certain block matrix. Hence, it is necessary to introduce two operators defined on block matrices:

$$\begin{aligned}\mathcal{P}_{\text{on}} \left(\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix} \right) &:= \begin{bmatrix}\mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}\end{bmatrix}, \\ \mathcal{P}_{\text{off}} \left(\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix} \right) &:= \begin{bmatrix}\mathbf{0} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{0}\end{bmatrix},\end{aligned} \tag{4.28}$$

for any matrices $\mathbf{A}_{11} \in \mathbb{R}^{n \times n}, \mathbf{A}_{12} \in \mathbb{R}^{n \times m}, \mathbf{A}_{21} \in \mathbb{R}^{m \times n}, \mathbf{A}_{22} \in \mathbb{R}^{m \times m}$.

According to the definitions of $\mathcal{P}_{\rm on}$ and $\mathcal{P}_{\rm off}$ in (4.28), when $\mathcal{P}_{\rm on}$ and $\mathcal{P}_{\rm off}$ are acting on the product of two block matrices $\mathbf{W}_1\mathbf{W}_2^\top$,

$$\begin{aligned}\mathcal{P}_{\rm on}(\mathbf{W}_1\mathbf{W}_2^\top) &= \mathcal{P}_{\rm on}\left(\begin{bmatrix}\mathbf{U}_1\mathbf{U}_2^\top & \mathbf{U}_1\mathbf{V}_2^\top\\ \mathbf{V}_1\mathbf{U}_2^\top & \mathbf{V}_1\mathbf{V}_2^\top\end{bmatrix}\right) = \begin{bmatrix}\mathbf{U}_1\mathbf{U}_2^\top & \mathbf{0}\\ \mathbf{0} & \mathbf{V}_1\mathbf{V}_2^\top\end{bmatrix} = \frac{\mathbf{W}_1\mathbf{W}_2^\top + \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_2^\top}{2},\\ \mathcal{P}_{\rm off}(\mathbf{W}_1\mathbf{W}_2^\top) &= \mathcal{P}_{\rm on}\left(\begin{bmatrix}\mathbf{U}_1\mathbf{U}_2^\top & \mathbf{U}_1\mathbf{V}_2^\top\\ \mathbf{V}_1\mathbf{U}_2^\top & \mathbf{V}_1\mathbf{V}_2^\top\end{bmatrix}\right) = \begin{bmatrix}\mathbf{0} & \mathbf{V}_1\mathbf{V}_2^\top\\ \mathbf{V}_1\mathbf{U}_2^\top & \mathbf{0}\end{bmatrix} = \frac{\mathbf{W}_1\mathbf{W}_2^\top - \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_2^\top}{2}.\end{aligned} \tag{4.29}$$

Here, to simplify the notations, for any $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{n\times r}$ and $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{R}^{m\times r}$, we define

$$\mathbf{W}_1 = \begin{bmatrix}\mathbf{U}_1\\ \mathbf{V}_1\end{bmatrix}, \qquad \widehat{\mathbf{W}}_1 = \begin{bmatrix}\mathbf{U}_1\\ -\mathbf{V}_1\end{bmatrix}, \qquad \mathbf{W}_2 = \begin{bmatrix}\mathbf{U}_2\\ \mathbf{V}_2\end{bmatrix}, \qquad \widehat{\mathbf{W}}_2 = \begin{bmatrix}\mathbf{U}_2\\ -\mathbf{V}_2\end{bmatrix}.$$

Now, we are ready to present the properties regarding the set $\mathcal{E}$ in Lemma 4.4.1 and Lemma 4.4.2, whose proofs are given in Appendix C.6 and Appendix C.7, respectively.

Lemma 4.4.1. *Let* $\mathbf{W} = \begin{bmatrix}\mathbf{U}^\top & \mathbf{V}^\top\end{bmatrix}^\top$ *with* $(\mathbf{U}, \mathbf{V}) \in \mathcal{E}$. *Then for every* $\mathbf{D} = \begin{bmatrix}\mathbf{D}_{\mathbf{U}}^\top & \mathbf{D}_{\mathbf{V}}^\top\end{bmatrix}^\top$ *of proper dimension, we have*

$$\|\mathcal{P}_{\rm on}(\mathbf{D}\mathbf{W}^\top)\|_F^2 = \|\mathcal{P}_{\rm off}(\mathbf{D}\mathbf{W}^\top)\|_F^2.$$

Lemma 4.4.2. *Let* $\mathbf{W}_1 = \begin{bmatrix}\mathbf{U}_1^\top & \mathbf{V}_1^\top\end{bmatrix}^\top$, $\mathbf{W}_2 = \begin{bmatrix}\mathbf{U}_2^\top & \mathbf{V}_2^\top\end{bmatrix}^\top$ *with* $(\mathbf{U}_1, \mathbf{V}_1), (\mathbf{U}_2, \mathbf{V}_2) \in \mathcal{E}$. *Then*

$$\|\mathcal{P}_{\rm on}(\mathbf{W}_1\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{W}_2^\top)\|_F^2 \le \|\mathcal{P}_{\rm off}(\mathbf{W}_1\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{W}_2^\top)\|_F^2.$$

4.4.5 Proof Idea: Connecting the Optimality Conditions

First observe that each $(\mathbf{U}^\star, \mathbf{V}^\star)$ in (4.22) is a global optimum for the factored program (we prove this in Appendix C.8):

Proposition 4.4.2. *Any* $(\mathbf{U}^\star, \mathbf{V}^\star)$ *in* (4.22) *is a global optimum of the factored program* $(\mathcal{F}_1)$*:*

$$g(\mathbf{U}^\star, \mathbf{V}^\star) \le g(\mathbf{U}, \mathbf{V}), \text{ for all } \mathbf{U} \in \mathbb{R}^{n\times r}, \mathbf{V} \in \mathbb{R}^{m\times r}.$$

However, due to non-convexity, only characterizing the global optima is not enough for the factored program to achieve the global convergence by many local-search algorithms. One should also eliminate the possibility of the existence of spurious local minima or degenerate saddles. For this purpose, we focus on the critical point set $\mathcal{X}$ and observe that any critical point $(\mathbf{U}, \mathbf{V}) \in \mathcal{X}$ of the factored problem satisfies the first part of the optimality condition (4.25):

$$\Xi(\mathbf{X})\mathbf{W} = \mathbf{0}$$

by constructing $\mathbf{W} = [\mathbf{U}^\top\ \mathbf{V}^\top]^\top$ and $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. If the critical point $(\mathbf{U}, \mathbf{V})$ additionally satisfies $\|\nabla f(\mathbf{U}\mathbf{V}^\top)\| \le \lambda$, then it corresponds to the global optimum $\mathbf{X}^\star = \mathbf{U}\mathbf{V}^\top$.

Therefore, it remains to study the additional critical points (which are introduced by the parameterization $\mathbf{X} = \psi(\mathbf{U}, \mathbf{V})$) that violate $\|\nabla f(\mathbf{U}\mathbf{V}^\top)\| \leq \lambda$. In fact, we intend to show the following: for any critical point $(\mathbf{U}, \mathbf{V})$, if $\mathbf{X}^\star \neq \mathbf{U}\mathbf{V}^\top$, we can find a direction $\mathbf{D}$, in which the Hessian $\nabla^2 g(\mathbf{U}, \mathbf{V})$ has a strictly negative curvature $[\nabla^2 g(\mathbf{U}, \mathbf{V})](\mathbf{D}, \mathbf{D}) < -\tau \|\mathbf{D}\|_F^2$ for some $\tau > 0$. Hence, every critical point $(\mathbf{U}, \mathbf{V})$ either corresponds to the global optimum $\mathbf{X}^\star$, or is a strict saddle point.

To gain more intuition, we take a closer look at the directional curvature of $g(\mathbf{U}, \mathbf{V})$ in some direction $\mathbf{D} = [\mathbf{D}_\mathbf{U}^\top \ \mathbf{D}_\mathbf{V}^\top]^\top$:

$$[\nabla^2 g(\mathbf{U}, \mathbf{V})](\mathbf{D}, \mathbf{D}) = \langle \Xi(\mathbf{X}), \mathbf{D}\mathbf{D}^\top \rangle + [\nabla^2 f(\mathbf{X})](\mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top, \mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top), \tag{4.30}$$

where the second term is always nonnegative by the convexity of f. The sign of the first term $\langle \Xi(\mathbf{X}), \mathbf{D}\mathbf{D}^\top \rangle$ depends on the positive semi-definiteness of $\Xi(\mathbf{X})$, which is related to the boundedness condition $\|\nabla f(\mathbf{X})\| \leq \lambda$ through the Schur complement theorem [115, A.5.5]:

$$\begin{aligned} & \Xi(\mathbf{X}) \succeq 0 \\ \iff & \lambda \mathbf{I} - \frac{1}{\lambda} \nabla f(\mathbf{X})^\top \nabla f(\mathbf{X}) \succeq 0 \\ \iff & \|\nabla f(\mathbf{X})\| \leq \lambda. \end{aligned}$$

Equivalently, whenever $\|\nabla f(\mathbf{X})\| > \lambda$, we have $\Xi(\mathbf{X}) \not\succeq 0$. Therefore, for those non-globally optimal critical points $(\mathbf{U}, \mathbf{V})$, it is possible to find a direction $\mathbf{D}$ such that the first term $\langle \Xi(\mathbf{X}), \mathbf{D}\mathbf{D}^\top \rangle$ is strictly negative. Inspired by the weighted PCA example, we choose $\mathbf{D}$ as the direction from the critical point $\mathbf{W} = \begin{bmatrix} \mathbf{U}^\top & \mathbf{V}^\top \end{bmatrix}^\top$ to the nearest globally optimal factor $\mathbf{W}^\star \mathbf{R}$ with $\mathbf{W}^\star = \begin{bmatrix} \mathbf{U}^{\star\top} & \mathbf{V}^{\star\top} \end{bmatrix}^\top$, i.e.,

$$\mathbf{D} = \mathbf{W} - \mathbf{W}^\star \mathbf{R},$$

where $\mathbf{R} = \arg\min_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top = \mathbf{I}_r} \|\mathbf{W} - \mathbf{W}^\star \mathbf{R}\|_F$. We will see that with this particular $\mathbf{D}$, the first term of (4.30) will be strictly negative while the second term retains small.

4.4.6 A Formal Proof of Theorem 4.4.1

The main argument involves choosing $\mathbf{D}$ as the direction from $\mathbf{W} = \begin{bmatrix} \mathbf{U}^\top & \mathbf{V}^\top \end{bmatrix}^\top$ to its nearest optimal factor: $\mathbf{D} = \mathbf{W} - \mathbf{W}^\star \mathbf{R}$ with $\mathbf{R} = \arg\min_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top = \mathbf{I}_r} \|\mathbf{W} - \mathbf{W}^\star \mathbf{R}\|_F$, and showing that the Hessian $\nabla^2 g(\mathbf{U}, \mathbf{V})$ has a strictly negative curvature in the direction of $\mathbf{D}$ whenever $\mathbf{W} \neq \mathbf{W}^\star$. To that end, we first introduce the following lemma (with its proof in Appendix C.9) connecting the distance $\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\|_F$ and the distance $\|(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star \mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\|_F$ (where $\mathbf{Q}\mathbf{Q}^\top$ is an orthogonal projector onto the $\mathrm{Span}(\mathbf{W})$).

Lemma 4.4.3. *Suppose the function $f(\mathbf{X})$ in $(\mathcal{P}_1)$ is restricted well-conditioned $(\mathcal{C})$. Let $\mathbf{W} = \begin{bmatrix} \mathbf{U}^\top & \mathbf{V}^\top \end{bmatrix}^\top$ with $(\mathbf{U}, \mathbf{V}) \in \mathcal{X}$, $\mathbf{W}^\star = \begin{bmatrix} \mathbf{U}^{\star\top} & \mathbf{V}^{\star\top} \end{bmatrix}^\top$ correspond to the global optimum of $(\mathcal{P}_1)$ and $\mathbf{Q}\mathbf{Q}^\top$ be the orthogonal*

projector onto $\mathrm{Range}(\mathbf{W})$. *Then*

$$\|(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\|_F \le 2\frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\|_F.$$

Proof of Theorem 4.4.1. Let $\mathbf{D} = \mathbf{W} - \mathbf{W}^\star\mathbf{R}$ with $\mathbf{R} = \arg\min_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{I}_r} \|\mathbf{W} - \mathbf{W}^\star\mathbf{R}\|_F$. Then

$$\begin{aligned}
&[\nabla^2 g(\mathbf{U},\mathbf{V})](\mathbf{D},\mathbf{D}) \\
&= \langle \Xi(\mathbf{X}), \mathbf{D}\mathbf{D}^\top\rangle + [\nabla^2 f(\mathbf{X})](\mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top, \mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top) \\
&\overset{①}{=} \langle \Xi(\mathbf{X}), \mathbf{W}^\star\mathbf{W}^{\star\top} - \mathbf{W}\mathbf{W}^\top\rangle + [\nabla^2 f(\mathbf{X})](\mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top, \mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top) \\
&\overset{②}{\le} \left\langle \Xi(\mathbf{X}) - \Xi(\mathbf{X}^\star), \mathbf{W}^\star\mathbf{W}^{\star\top} - \mathbf{W}\mathbf{W}^\top\right\rangle + [\nabla^2 f(\mathbf{X})](\mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top, \mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top) \\
&= \left\langle \begin{bmatrix} \lambda\mathbf{I} & \nabla f(\mathbf{X}) \\ \nabla f(\mathbf{X})^\top & \lambda\mathbf{I}\end{bmatrix} - \begin{bmatrix} \lambda\mathbf{I} & \nabla f(\mathbf{X}^\star) \\ \nabla f(\mathbf{X}^\star)^\top & \lambda\mathbf{I}\end{bmatrix}, \mathbf{W}^\star\mathbf{W}^{\star\top} - \mathbf{W}\mathbf{W}^\top\right\rangle + [\nabla^2 f(\mathbf{X})](\mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top, \mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top) \\
&\overset{③}{=} \left\langle \begin{bmatrix} \mathbf{0} & * \\ *^\top & \mathbf{0}\end{bmatrix}, \mathbf{W}^\star\mathbf{W}^{\star\top} - \mathbf{W}\mathbf{W}^\top\right\rangle + [\nabla^2 f(\mathbf{X})](\mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top, \mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top) \\
&= -2\int_0^1 [\nabla^2 f(\mathbf{X}^\star + t(\mathbf{X}-\mathbf{X}^\star))](\mathbf{X}-\mathbf{X}^\star, \mathbf{X}-\mathbf{X}^\star)dt + [\nabla^2 f(\mathbf{X})](\mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top, \mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top)
\end{aligned}$$

where ① follows from $\nabla g(\mathbf{U},\mathbf{V}) = \Xi(\mathbf{X})\mathbf{W} = \mathbf{0}$ and (4.26). For ②, we note that $\langle \Xi(\mathbf{X}^\star), \mathbf{W}^\star\mathbf{W}^{\star\top} - \mathbf{W}\mathbf{W}^\top\rangle \le 0$ since $\Xi(\mathbf{X}^\star)\mathbf{W}^\star = \mathbf{0}$ in (4.25) and $\Xi(\mathbf{X}^\star) \succeq 0$ by the optimality condition. In ③, we use

$$* = (\int_0^1 [\nabla^2 f(\mathbf{X}^\star + t(\mathbf{X}-\mathbf{X}^\star))](\mathbf{X}-\mathbf{X}^\star)dt)$$

for convenience and then ③ follows from the Taylor's Theorem for vector-valued functions [148, Eq. (2.5) in Theorem 2.1]:

$$\nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^\star) = \int_0^1 [\nabla^2 f(\mathbf{X}^\star + t(\mathbf{X}-\mathbf{X}^\star))](\mathbf{X}-\mathbf{X}^\star)dt.$$

Now, we continue the argument:

$$\begin{aligned}
&[\nabla^2 g(\mathbf{U},\mathbf{V})](\mathbf{D},\mathbf{D}) \\
&\le -2\int_0^1 [\nabla^2 f(\mathbf{X}^\star + t(\mathbf{X}-\mathbf{X}^\star))](\mathbf{X}-\mathbf{X}^\star, \mathbf{X}-\mathbf{X}^\star)dt + [\nabla^2 f(\mathbf{X})](\mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top, \mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top) \\
&\overset{④}{\le} -2\alpha\|\mathbf{X}^\star - \mathbf{X}\|_F^2 + \beta\|\mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top\|_F^2, \\
&\overset{⑤}{\le} -0.5\alpha\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\|_F^2 + 2\beta(\|\mathbf{D}_\mathbf{U}\mathbf{V}^\top\|_F^2 + \|\mathbf{U}\mathbf{D}_\mathbf{V}^\top\|_F^2) \\
&\overset{⑥}{=} -0.5\alpha\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\|_F^2 + \beta\|\mathbf{D}\mathbf{W}^\top\|_F^2 \\
&\overset{⑦}{\le} \left[-0.5\alpha + \beta/8 + 4.208\beta\left(\frac{\beta-\alpha}{\beta+\alpha}\right)^2\right]\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\|_F^2 \\
&\overset{⑧}{\le} -0.06\alpha\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\|_F^2
\end{aligned}$$

$$\overset{⑨}{\leq} \begin{cases} -0.06\alpha \min\left\{\rho^2(\mathbf{W}), \rho^2(\mathbf{W}^\star)\right\} \|\mathbf{D}\|_F^2, & \text{By Lemma 4.3.2 when } r > r^\star \\ -0.0495\alpha\rho^2(\mathbf{W}^\star)\|\mathbf{D}\|_F^2, & \text{By Lemma 4.3.3 when } r = r^\star \\ -0.06\alpha\rho^2(\mathbf{W}^\star)\|\mathbf{D}\|_F^2, & \text{When } \mathbf{W} = \mathbf{0} \end{cases}$$

where ④ uses the restricted well-conditionedness ($\mathcal{C}$) since $\mathrm{rank}(\mathbf{X}^\star + t(\mathbf{X} - \mathbf{X}^\star)) \leq 2r$, $\mathrm{rank}(\mathbf{X} - \mathbf{X}^\star) \leq 4r$ and $\mathrm{rank}(\mathbf{D}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{D}_\mathbf{V}^\top) \leq 4r$. ⑤ comes from Lemma 4.4.2 and the fact $\|\mathbf{A} + \mathbf{B}\|_F^2 \leq 2(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2)$. ⑥ follows from Lemma 4.4.1. ⑦ first uses Lemma 4.3.4 to bound $\|\mathbf{D}\mathbf{W}^\top\|_F^2 = \|(\mathbf{W} - \mathbf{W}^\star\mathbf{R})\mathbf{W}^\top\|_F^2$ since $\mathbf{W}^\top\mathbf{W}^\star \succeq 0$ and then uses Lemma 4.4.3 to further bound $\|(\mathbf{W}^\star - \mathbf{W})\mathbf{Q}\mathbf{Q}^\top\|_F^2$. ⑧ holds when $\beta/\alpha \leq 1.5$. ⑨ uses the similar argument as in the proof of Theorem 4.3.1 to relate the lifted distance and factored distance. Particularly, three possible cases are considered: (i) $r > r^\star$; (ii) $r = r^\star$; (iii) $\mathbf{W} = \mathbf{0}$. We apply Lemma 4.3.2 to Case (i) and Lemma 4.3.3 to Case (ii). For the third case that $\mathbf{W} = \mathbf{0}$, we obtain from ⑧ that

$$[\nabla^2 g(\mathbf{U}, \mathbf{V})](\mathbf{D}, \mathbf{D}) \leq -0.06\alpha\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F^2 \leq -0.06\alpha\rho(\mathbf{W}^\star)^2\|\mathbf{W}^\star\|_F^2 = -0.06\alpha\rho(\mathbf{W}^\star)^2\|\mathbf{D}\|_F^2,$$

where the last equality follows from $\mathbf{D} = \mathbf{0} - \mathbf{W}^\star\mathbf{R} = -\mathbf{W}^\star\mathbf{R}$ because $\mathbf{W} = \mathbf{0}$.

The final result follows from the the definition of $\mathbf{U}^\star, \mathbf{V}^\star$ in (4.22):

$$\mathbf{W}^\star = \begin{bmatrix} \mathbf{P}^\star\sqrt{\mathbf{\Sigma}^\star}\mathbf{R} \\ \mathbf{Q}^\star\sqrt{\mathbf{\Sigma}^\star}\mathbf{R} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^\star/\sqrt{2} \\ \mathbf{Q}^\star/\sqrt{2} \end{bmatrix} \left(\sqrt{2\mathbf{\Sigma}^\star}\right)\mathbf{R},$$

which implies $\sigma_\ell(\mathbf{W}^\star) = \sqrt{2\sigma_\ell(\mathbf{X}^\star)}$. □

4.5 Conclusion

In this work, we considered two popular minimization problems: the minimization of a general convex function $f(\mathbf{X})$ with the domain being positive semi-definite matrices; the minimization of a general convex function $f(\mathbf{X})$ regularized by the matrix nuclear norm $\|\mathbf{X}\|_*$ with the domain being general matrices. To improve the computational efficiency, we applied the Burer-Monteiro re-parameterization and showed that, as long as the convex function $f(\mathbf{X})$ is (restricted) well-conditioned, the resulting factored problems have the following properties: each critical point either corresponds to a global optimum of the original convex programs, or is a strict saddle where the Hessian matrix has a strictly negative eigenvalue. Such a benign landscape then allows many iterative optimization methods to escape from all the saddle points and converge to a global optimum with even random initializations.

CHAPTER 5

GLOBAL OPTIMALITY IN LOW-RANK MATRIX OPTIMIZATION

This work[11] considers the minimization of a general objective function $f(\mathbf{X})$ over the set of rectangular $n \times m$ matrices that have rank at most r. To reduce the computational burden, we factorize the variable $\mathbf{X}$ into a product of two smaller matrices and optimize over these two matrices instead of $\mathbf{X}$. Despite the resulting nonconvexity, recent studies in matrix completion and sensing have shown that the factored problem has no spurious local minima and obeys the so-called strict saddle property (the function has a directional negative curvature at all critical points but local minima). We analyze the global geometry for a general and yet well-conditioned objective function $f(\mathbf{X})$ whose restricted strong convexity and restricted strong smoothness constants are comparable. In particular, we show that the reformulated objective function has no spurious local minima and obeys the strict saddle property. These geometric properties imply that a number of iterative optimization algorithms (such as gradient descent) can provably solve the factored problem with global convergence.

5.1 Introduction

Consider the minimization of a general objective function $f(\mathbf{X})$ over all low-rank $n \times m$ matrices:

$$\begin{aligned} &\underset{\mathbf{X}\in\mathbb{R}^{n\times m}}{\text{minimize}}\ f(\mathbf{X}) \\ &\text{subject to } \operatorname{rank}(\mathbf{X}) \le r, \end{aligned} \tag{5.1}$$

where the objective function $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ is smooth. Low-rank matrix optimizations of the form (5.1) appear in a wide variety of applications, including quantum tomography [149, 150], collaborative filtering [117, 151], sensor localization [118], low-rank matrix recovery from compressive measurements [65, 152], and matrix completion [153, 154]. Due to the rank constraint, however, low-rank matrix optimizations of the form (5.1) are highly nonconvex and computationally NP-hard in general [155] even if f itself is convex. In order to deal with the rank constraint and to find a low-rank solution, the nuclear norm is widely used in matrix inverse problems [65, 156] arising in machine learning [157], signal processing [124], and control [158]. Although nuclear norm minimization enjoys strong statistical guarantees [153], its computational complexity is very high (as most algorithms require performing an expensive singular value decomposition (SVD) in each iteration), prohibiting it from scaling to practical problems.

To relieve the computational bottleneck and provide an alternative way of dealing with the rank constraint, recent studies propose to factorize the variable into the Burer-Monteiro type decomposition [119, 159] with $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, and optimize over the $n \times r$ and $m \times r$ matrices $\mathbf{U}$ and $\mathbf{V}$. With this parameterization of $\mathbf{X}$, we can recast (5.1) into the

[11] This is a joint work with Zhihui Zhu, Gongguo Tang and Michael B. Wakin [8].

following program:

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \; h(\mathbf{U},\mathbf{V}) := f(\mathbf{U}\mathbf{V}^\top). \tag{5.2}$$

The bilinear nature of the parameterization renders the objective function of (5.2) nonconvex even when $f(\mathbf{X})$ is a convex function. Hence, the objective function in (5.2) can potentially have spurious local minima (i.e., local minimizers that are not global minimizers) or "bad" saddle points that prevent a number of iterative algorithms from converging to the global solution. By analyzing the landscape of nonconvex functions, several recent works have shown that the factored objective function $h(\mathbf{U},\mathbf{V})$ in certain matrix inverse problems has no spurious local minima [100, 106, 125].

We generalize this line of work by focusing on a general objective function $f(\mathbf{X})$ in the optimization (5.1), not necessarily a quadratic loss function coming from a matrix inverse problem. By focusing on a general objective function, we attempt to provide a unifying framework for low-rank matrix optimizations with the factorization approach. We provide a geometric analysis for the factored program (5.2) and show that, under certain conditions on $f(\mathbf{X})$, all critical points of the objective function $h(\mathbf{U},\mathbf{V})$ are well-behaved. Our characterization of the geometry of the objective function ensures that a number of iterative optimization algorithms converge to a global minimum.

5.1.1 Summary of Results

The purpose of this work is to analyze the geometry of the factored problem $h(\mathbf{U},\mathbf{V})$ in (5.2). In particular, we attempt to understand the behavior of all of the critical points of the objective function in the reformulated problem (5.2).

Before presenting our main results, we lay out the necessary assumptions on the objective function $f(\mathbf{X})$. As is known, without any assumptions on the problem, even minimizing traditional quadratic objective functions is challenging. For this purpose, we focus on the model where $f(\mathbf{X})$ is $(2r,4r)$-restricted strongly convex and smooth, i.e., for any $n\times m$ matrices $\mathbf{X},\mathbf{G}$ with $\mathrm{rank}(\mathbf{X})\le 2r$ and $\mathrm{rank}(\mathbf{G})\le 4r$, the Hessian of $f(\mathbf{X})$ satisfies

$$\alpha\,\|\mathbf{G}\|_F^2 \le [\nabla^2 f(\mathbf{X})](\mathbf{G},\mathbf{G}) \le \beta\,\|\mathbf{G}\|_F^2 \tag{5.3}$$

for some positive α and β. A similar assumption is also utilized in [109, Conditions 5.3 and 5.4]. With this assumption on $f(\mathbf{X})$, we summarize our main results in the following informal theorem.

Theorem 5.1.1. (informal) *Suppose the function $f(\mathbf{X})$ satisfies the $(2r,4r)$-restricted strong convexity and smoothness condition* (5.3) *and has a critical point $\mathbf{X}^\star\in\mathbb{R}^{n\times m}$ with $\mathrm{rank}(\mathbf{X}^\star)=r^\star\le r$. Then the factored objective function $h(\mathbf{U},\mathbf{V})$ (with an additional regularizer, see Theorem 5.3.1) in* (5.2) *has no spurious local minima and obeys the strict saddle property (see Definition 5.2.3 in Section 5.2).*

Remark 5.1.1. As guaranteed by Proposition 5.3.1 (in Section 5.3), the $(2r,4r)$-restricted strong convexity and smoothness property (5.3) ensures that $\mathbf{X}^\star$ is the unique global minimum of (5.1). Theorem 5.1.1 then implies

that we can recover the rank-$r^\star$ global minimizer $\mathbf{X}^\star$ of (5.1) by many iterative algorithms (such as the trust region method [160] and stochastic gradient descent [114]) even from a random initialization. This is because 1) as guaranteed by Theorem 5.2.1, the strict saddle property ensures local search algorithms converge to a local minimum, and 2) there are no spurious local minima.

Remark 5.1.2. Since our main result only requires the $(2r, 4r)$-restricted strong convexity and smoothness property (5.3), aside from low-rank matrix recovery [156], it can also be applied to many other low-rank matrix optimization problems [161] which do not necessarily involve quadratic loss functions. Typical examples include robust PCA [162, 163], 1-bit matrix completion [132, 164] and Poisson principal component analysis (PCA) [165].

Remark 5.1.3. Similar results on positive semi-definite (PSD) matrix optimization problems (but without the rank constraint) with generic objective functions were obtained in [6]. We note that one cannot directly apply the results in [6] to the optimization (5.1) when the matrices under consideration are nonsymmetric or rectangular, even if we ignore the rank constraint. One could attempt to convert minimizing $f(\mathbf{X})$ over general $n \times m$ matrices into minimizing $q(\mathbf{Z})$ over the cone of PSD matrices of size $(m+n) \times (m+n)$, where $\mathbf{X}$ and $\mathbf{X}^\top$ form the upper right and lower left blocks of $\mathbf{Z}$. The problem with this transformation, however, is that $q(\mathbf{Z})$ will no longer satisfy the same properties as $f(\mathbf{X})$, in particular the restricted strong convexity and smoothness condition (5.3) which is a key assumption utilized in [6]. For this reason, one cannot apply the results for the PSD optimization in [6] directly to our problem. In terms of the proof techniques, although the generalization from the PSD case might not seem technically challenging at first sight, quite a few technical difficulties had to be overcome to develop the theory for the general case in this work. In fact, the non-triviality of extending to the nonsymmetric case is also highlighted in [102, 106].

5.1.2 Related Works

Compared with the original program (5.1), the factored form (5.2) typically involves many fewer variables (or variables with much smaller size) and can be efficiently solved by simple but powerful methods (such as gradient descent [113, 114], the trust region method [97], and alternating methods [166]) for large-scale settings, though it is nonconvex. In recent years, tremendous effort has been devoted to analyzing nonconvex optimizations by exploiting the geometry of the corresponding objective functions. These works can be separated into two types based on whether the geometry is analysed locally or globally. One type of work analyzes the behavior of the objective function in a small neighborhood containing the global optimum and requires a good initialization that is close enough to a global minimum. Problems such as phase retrieval [167], matrix sensing [102], and semi-definite optimization [103] have been studied.

Another type of work attempts to analyze the landscape of the objective function and show that it obeys the strict saddle property. If this particular property holds, then simple algorithms such as gradient descent and the trust region method are guaranteed to converge to a local minimum from a random initialization [113, 114, 168] rather than

requiring a good guess. We approach low-rank matrix optimization with general objective functions (5.1) via a similar geometric characterization. Similar geometric results are known for a number of problems including complete dictionary learning [168], phase retrieval [160], orthogonal tensor decomposition [114], and matrix inverse problems [6, 100, 125]. Empirical evidence also supports using the factorization approach for estimating a low-rank PSD matrix from a set of rank-one measurements corrupted by arbitrary outliers [120] and for recovering a dynamically evolving low-rank matrix from incomplete observations [102, 169].

Our work is most closely related to certain recent works in low-rank matrix optimization. Bhojanapalli et al. [125] showed that the low-rank, PSD matrix sensing problem has no spurious local minima and obeys the strict saddle property. Similar results were exploited for PSD matrix completion [100], PSD matrix factorization [104] and low-rank, PSD matrix optimization problems with generic objective functions [6]. Our work extends this line of analysis to general low-rank matrix (not necessary PSD or even square) optimization problems. Another closely related work considers the low-rank, non-square matrix sensing problem and matrix completion with the factorization approach [93, 101, 106]. We note that our general objective function framework includes the low-rank matrix sensing problem as a special case (see Section 5.3.3). Furthermore, our result covers both over-parameterization where $r > r^\star$ and exact parameterization where $r = r^\star$. Wang et al. [109] also considered the factored low-rank matrix minimization problem with a general objective function which satisfies the restricted strong convexity and smoothness condition. Their algorithms require good initializations for global convergence since they characterized only the local landscapes around the global optima. By categorizing the behavior of all the critical points, our work differs from [109] in that we instead characterize the global landscape of the factored objective function.

This chapter continues in Section 5.2 with formal definitions for strict saddles and the strict saddle property. We present the main results and their implications in matrix sensing, weighted low-rank approximation, and 1-bit matrix completion in Section 5.3. The proof of our main results is given in Section 5.4. We conclude the chapter in Section 5.6.

5.2 Preliminaries

5.2.1 Notation

To begin, we first briefly introduce some notation used throughout the chapter. The symbols $\mathbf{I}$ and $\mathbf{0}$ respectively represent the identity matrix and zero matrix with appropriate sizes. The set of $r \times r$ orthonormal matrices is denoted by $\mathcal{O}_r := \{\mathbf{R} \in \mathbb{R}^{r\times r} : \mathbf{R}^\top\mathbf{R} = \mathbf{I}\}$. If a function $h(\mathbf{U}, \mathbf{V})$ has two arguments, $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{V} \in \mathbb{R}^{m\times r}$, we occasionally use the notation $h(\mathbf{W})$ when we put these two arguments into a new one as $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$. For a scalar function $f(\mathbf{Z})$ with a matrix variable $\mathbf{Z} \in \mathbb{R}^{n\times m}$, its gradient is an $n \times m$ matrix whose (i,j)-th entry is $[\nabla f(\mathbf{Z})]_{ij} = \frac{\partial f(\mathbf{Z})}{\partial Z_{ij}}$ for all $i \in [n], j \in [m]$. Here $[n] = \{1, 2, \ldots, n\}$ for any $n \in \mathbb{N}$ and Z_{ij} is the (i,j)-th entry of the matrix $\mathbf{Z}$. The Hessian of $f(\mathbf{Z})$ can be viewed as an $nm \times nm$ matrix $[\nabla^2 f(\mathbf{Z})]_{ij} = \frac{\partial^2 f(\mathbf{Z})}{\partial z_i \partial z_j}$ for all $i, j \in [nm]$,

where z_i is the i-th entry of the vectorization of $\mathbf{Z}$. An alternative way to represent the Hessian is by a bilinear form defined via $[\nabla^2 f(\mathbf{Z})](\mathbf{A}, \mathbf{B}) = \sum_{i,j,k,l} \frac{\partial^2 f(\mathbf{Z})}{\partial Z_{ij} \partial Z_{kl}} A_{ij} B_{kl}$ for any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times m}$. The bilinear form for the Hessian is widely utilized through the chapter.

5.2.2 Strict Saddle Property

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable objective function. We begin with the notion of strict saddles and the strict saddle property.

Definition 5.2.1 (Critical points)**.** *We say* $\mathbf{x}$ *a critical point if the gradient at* $\mathbf{x}$ *vanishes, i.e.,* $\nabla h(\mathbf{x}) = \mathbf{0}$.

Definition 5.2.2 (Strict saddles)**.** *A critical point* $\mathbf{x}$ *is a strict saddle if the Hessian matrix evaluated at this point has a strictly negative eigenvalue, i.e.,* $\lambda_{\min}(\nabla^2 h(\mathbf{x})) < 0$.

Definition 5.2.3 (Strict saddle property [114])**.** *A twice differentiable function satisfies the strict saddle property if each critical point either corresponds to a local minimum or is a strict saddle.*

Intuitively, the strict saddle property requires a function to have a directional negative curvature at all critical points but local minima. This property allows a number of iterative algorithms such as noisy gradient descent [114] and the trust region method [170] to further decrease the function value at all the strict saddles and thus converge to a local minimum.

Theorem 5.2.1. *[97, 113, 114] (informal) For a twice continuously differentiable objective function satisfying the strict saddle property, a number of iterative optimization algorithms (such as gradient descent and the the trust region method) can find a local minimum.*

5.3 Problem Formulation and Main Results

5.3.1 Problem Formulation

This work considers the problem (5.1) of minimizing a general function $f(\mathbf{X})$ (over the set of low-rank matrices) which is assumed to have a low-rank critical point $\mathbf{X}^\star$ with $\mathrm{rank}(\mathbf{X}^\star) = r^\star \leq r$ such that $\nabla f(\mathbf{X}^\star) = \mathbf{0}$. Because of the restricted strong convexity and smoothness condition (5.3), the following result establishes that if $f(\mathbf{X})$ has a critical point $\mathbf{X}^\star$ with $\mathrm{rank}(\mathbf{X}^\star) \leq r$, then it is the unique global minimum of (5.1).

Proposition 5.3.1. *Suppose* $f(\mathbf{X})$ *satisfies the* $(2r, 4r)$*-restricted strong convexity and smoothness condition* (5.3) *with positive* α *and* β*. Assume* $\mathbf{X}^\star$ *is a critical point of* $f(\mathbf{X})$ *with* $\mathrm{rank}(\mathbf{X}^\star) = r^\star \leq r$*. Then* $\mathbf{X}^\star$ *is the global minimum of* (5.1)*, i.e.,*

$$f(\mathbf{X}^\star) \leq f(\mathbf{X}),\ \forall\ \mathbf{X} \in \mathbb{R}^{n \times m}, \mathrm{rank}(\mathbf{X}) \leq r$$

and the equality holds only at $\mathbf{X} = \mathbf{X}^\star$.

Proof of Proposition 5.3.1. First note that if $\mathbf{X}^\star$ is a critical point of $f(\mathbf{X})$, then

$$\nabla f(\mathbf{X}^\star) = \mathbf{0}.$$

Now for any $\mathbf{X} \in \mathbb{R}^{n\times m}$ with $\mathrm{rank}(\mathbf{X}) \le r$, the second order Taylor expansion gives

$$f(\mathbf{X}) = f(\mathbf{X}^\star) + \langle \nabla f(\mathbf{X}^\star), \mathbf{X} - \mathbf{X}^\star \rangle + \frac{1}{2}[\nabla^2 f(\widetilde{\mathbf{X}})](\mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star),$$

where $\widetilde{\mathbf{X}} = t\mathbf{X}^\star + (1-t)\mathbf{X}$ for some $t \in [0,1]$. This Taylor expansion together with $\nabla f(\mathbf{X}^\star) = \mathbf{0}$ and (5.3) (both $\widetilde{\mathbf{X}}$ and $\mathbf{X}' - \mathbf{X}^\star$ have rank at most $2r$) gives

$$\begin{aligned} f(\mathbf{X}) - f(\mathbf{X}^\star) &= \frac{1}{2}[\nabla^2 f(\widetilde{\mathbf{X}})](\mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star) \\ &\ge \frac{\alpha}{2}\|\mathbf{X} - \mathbf{X}^\star\|_F^2. \end{aligned}$$

□

With this, in the sequel, we use $\mathbf{X}^\star$ to denote the global minimum of (5.1) (i.e., the low-rank critical point of $f(\mathbf{X})$), unless stated otherwise. We note that the assumption of the existence of a low-rank critical point $\mathbf{X}^\star$ is very mild and holds in many matrix inverse problems [65, 153], where the unknown matrix to be recovered is a critical point of f. We factorize the variable $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{n\times r}, \mathbf{V} \in \mathbb{R}^{m\times r}$ and transform (5.1) into its factored counterpart (5.2). Throughout the chapter, $\mathbf{X}$, $\mathbf{W}$ and $\widehat{\mathbf{W}}$ are matrices depending on $\mathbf{U}$ and $\mathbf{V}$:

$$\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}, \ \widehat{\mathbf{W}} = \begin{bmatrix}\mathbf{U}\\ -\mathbf{V}\end{bmatrix}, \ \mathbf{X} = \mathbf{U}\mathbf{V}^\top.$$

Although the new variable $\mathbf{W}$ has much smaller size than $\mathbf{X}$ when $r \ll \min\{n, m\}$, the objective function in the factored problem (5.2) may have a much more complicated landscape due to the bilinear form about $\mathbf{U}$ and $\mathbf{V}$. The reformulated objective function $h(\mathbf{U}, \mathbf{V})$ could introduce spurious local minima or degenerate saddle points even when $f(\mathbf{X})$ is convex. Our goal is to guarantee that this does not happen.

Let $\mathbf{X}^\star = \mathbf{Q}_{\mathbf{U}^\star}\mathbf{\Sigma}^\star\mathbf{Q}_{\mathbf{V}^\star}^\top$ denote an SVD of $\mathbf{X}^\star$, where $\mathbf{Q}_{\mathbf{U}^\star} \in \mathbb{R}^{n\times r}$ and $\mathbf{Q}_{\mathbf{V}^\star} \in \mathbb{R}^{m\times r}$ are orthonormal matrices of appropriate sizes, and $\mathbf{\Sigma}^\star \in \mathbb{R}^{r\times r}$ is a diagonal matrix with non-negative diagonal (but with some zeros on the diagonal if $r > r^\star = \mathrm{rank}(\mathbf{X}^\star)$). We denote

$$\mathbf{U}^\star = \mathbf{Q}_{\mathbf{U}^\star}\mathbf{\Sigma}^{\star 1/2}, \quad \mathbf{V}^\star = \mathbf{Q}_{\mathbf{V}^\star}\mathbf{\Sigma}^{\star 1/2},$$

where $\mathbf{X}^\star = \mathbf{U}^\star\mathbf{V}^{\star\top}$ forms a balanced factorization of $\mathbf{X}^\star$ since $\mathbf{U}^\star$ and $\mathbf{V}^\star$ have the same singular values. Throughout the chapter, we utilize the following two ways to stack $\mathbf{U}^\star$ and $\mathbf{V}^\star$ together:

$$\mathbf{W}^\star = \begin{bmatrix}\mathbf{U}^\star\\ \mathbf{V}^\star\end{bmatrix}, \quad \widehat{\mathbf{W}}^\star = \begin{bmatrix}\mathbf{U}^\star\\ -\mathbf{V}^\star\end{bmatrix}.$$

Before moving on, we note that for any solution $(\mathbf{U}, \mathbf{V})$ to (5.2), $(\mathbf{U}\boldsymbol{\Psi}, \mathbf{V}\boldsymbol{\Phi})$ is also a solution to (5.2) for any $\boldsymbol{\Psi}, \boldsymbol{\Phi} \in \mathbb{R}^{r\times r}$ such that $\mathbf{U}\boldsymbol{\Psi}\boldsymbol{\Phi}^\top\mathbf{V}^\top = \mathbf{U}\mathbf{V}^\top$. In order to address this ambiguity (i.e., to reduce the search space of $\mathbf{W}$ for (5.2)), we utilize the trick in [102, 106, 109] by introducing a regularizer

$$g(\mathbf{U}, \mathbf{V}) = \frac{\mu}{4}\left\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right\|_F^2 \tag{5.4}$$

and solving the following problem

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}, \mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \ \rho(\mathbf{U}, \mathbf{V}) := f(\mathbf{U}\mathbf{V}^\top) + g(\mathbf{U}, \mathbf{V}), \tag{5.5}$$

where $\mu > 0$ controls the weight for the term $\left\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right\|_F^2$, which will be discussed soon.

We remark that $\mathbf{W}^\star$ is still a global minimizer of the factored problem (5.5) since $f(\mathbf{X})$ achieves its global minimum over the low-rank set of matrices at $\mathbf{X}^\star$ and $g(\mathbf{W})$ also achieves its global minimum at $\mathbf{W}^\star$. The regularizer $g(\mathbf{W})$ is applied to force the difference between the two Gram matrices of $\mathbf{U}$ and $\mathbf{V}$ to be as small as possible. The global minimum of $g(\mathbf{W})$ is 0, which is achieved when $\mathbf{U}$ and $\mathbf{V}$ have the same Gram matrices, i.e., when $\mathbf{W}$ belongs to

$$\mathcal{E} := \left\{\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V} = \mathbf{0}\right\}. \tag{5.6}$$

Informally, we can view (5.5) as finding a point from $\mathcal{E}$ that also minimizes $f(\mathbf{U}\mathbf{V}^\top)$. This is formally established in Theorem 5.3.1.

5.3.2 Main Results

Our main argument is that, under certain conditions on $f(\mathbf{X})$, the objective function $\rho(\mathbf{W})$ has no spurious local minima and satisfies the strict saddle property. This is equivalent to categorizing all the critical points into two types: 1) the global minima which correspond to the global solution of the original convex problem (5.1) and 2) strict saddles such that the Hessian matrix $\nabla^2\rho(\mathbf{W})$ evaluated at these points has a strictly negative eigenvalue. We formally establish this in the following theorem, whose proof is given in the next section.

Theorem 5.3.1. *For any $\mu > 0$, each critical point $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ of $\rho(\mathbf{W})$ defined in* (5.5) *satisfies*

$$\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V} = \mathbf{0}. \tag{5.7}$$

Furthermore, suppose that the function $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (5.3) *with positive constants α and β satisfying $\frac{\beta}{\alpha} \le 1.5$ and that the function $f(\mathbf{X})$ has a critical point $\mathbf{X}^\star \in \mathbb{R}^{n\times m}$ with $\operatorname{rank}(\mathbf{X}^\star) = r^\star \le r$. Set $\mu \le \frac{\alpha}{16}$ for the factored problem* (5.5)*. Then $\rho(\mathbf{W})$ has no spurious local*

minima, i.e., any local minimum of $\rho(\mathbf{W})$ is a global minimum corresponding to the global solution of the original problem (5.1)*: $\mathbf{U}\mathbf{V}^\top = \mathbf{X}^\star$. In addition, $\rho(\mathbf{W})$ obeys the strict saddle property that any critical point not being a local minimum is a strict saddle with*

$$\lambda_{\min}\left(\nabla^2\left(\rho(\mathbf{W})\right)\right) \leq \begin{cases} -0.08\alpha\sigma_r(\mathbf{X}^\star), & r = r^\star \\ -0.05\alpha \cdot \min\left\{\sigma^2_{r^c}(\mathbf{W}), 2\sigma_{r^\star}(\mathbf{X}^\star)\right\}, & r > r^\star \\ -0.1\alpha\sigma_{r^\star}(\mathbf{X}^\star), & r^c = 0, \end{cases} \tag{5.8}$$

where $r^c \leq r$ is the rank of $\mathbf{W}$, $\lambda_{\min}(\cdot)$ represents the smallest eigenvalue, and $\sigma_\ell(\cdot)$ denotes the ℓ-th largest singular value.

Remark 5.3.1. Equation (5.7) shows that any critical point $\mathbf{W}$ belongs to $\mathcal{E}$ for the objective function in the factored problem (5.5) with any positive μ. This demonstrates the reason for adding the regularizer $g(\mathbf{U}, \mathbf{V})$. Thus, any iterative optimization algorithm converging to some critical point of $\rho(\mathbf{W})$ results in a solution within $\mathcal{E}$. Furthermore, the strict saddle property along with the lack of spurious local minima ensures that a number of iterative optimization algorithms find the global minimum.

Remark 5.3.2. For any critical point $\mathbf{W} \in \mathbb{R}^{(n+m)\times r}$ that is not a local minimum, the right hand side of (5.8) is strictly negative, implying that $\mathbf{W}$ is a strict saddle. We also note that Theorem 5.3.1 not only covers exact parameterization where $r = r^\star$, but also includes the over-parameterization case where $r > r^\star$.

Remark 5.3.3. The constants appearing in Theorem 5.3.1 are not optimized. We use $\mu \leq \frac{1}{16}\alpha$ simply to include $\mu = \frac{1}{16}$ which is utilized for the matrix sensing problem in [102]. If the ratio between the restricted strong convexity and smoothness constants $\frac{\beta}{\alpha} \leq 1.4$, then we can show that $\rho(\mathbf{W})$ has no spurious local minima and obeys the strict saddle property for any $\mu \leq \frac{1}{4}\alpha$ (where $\mu = \frac{1}{4}$ is utilized for the matrix sensing problem in [106]). In all cases, a smaller μ yields a more negative constant in (5.8); see Section 5.4 for more discussion on this. This implies that when the restricted strong convexity constant α is not provided a priori, one can always choose a small μ to ensure the strict saddle property holds, and hence guarantee the global convergence of many iterative optimization algorithms.

The constant 1.5 for the dynamic range $\frac{\beta}{\alpha}$ in Theorem 5.3.1 is also not optimized and it is possible to slightly relax this constraint with more sophisticated analysis. However, the following example involving weighted symmetric matrix factorization implies that the room for improving this constant is rather limited. Let

$$\mathbf{\Omega} = \begin{bmatrix} \sqrt{1+a} & 1 \\ 1 & \sqrt{1+a} \end{bmatrix}$$

for some $a \geq 0$,

$$\mathbf{X}^\star = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \text{ and } \mathbf{U} = \begin{bmatrix} x \\ y \end{bmatrix}.$$

Now consider the following weighted low-rank matrix factorization:

$$h(\mathbf{U}) = \frac{1}{2}\|\mathbf{\Omega} \odot (\mathbf{U}\mathbf{U}^\top - \mathbf{X}^\star)\|_F^2 = \frac{1+a}{2}\left(x^2-1\right)^2 + \frac{1+a}{2}\left(y^2-1\right)^2 + (xy-1)^2, \tag{5.9}$$

whose gradient $\nabla h(\mathbf{U})$ and Hessian $\nabla^2 h(\mathbf{U})$ are given by:

$$\nabla h(\mathbf{U}) = 2\begin{bmatrix}(a+1)\left(x^2-1\right)x + y(xy-1)\\ (a+1)\left(y^2-1\right)y + x(xy-1)\end{bmatrix},$$

and

$$\nabla^2 h(\mathbf{U}) = 2\begin{bmatrix}y^2 + \left(3x^2-1\right)(a+1) & 2xy-1\\ 2xy-1 & x^2 + \left(3y^2-1\right)(a+1)\end{bmatrix}.$$

Then,

$$\mathbf{U} = \begin{bmatrix}\sqrt{\frac{a}{a+2}}\\ -\sqrt{\frac{a}{a+2}}\end{bmatrix}$$

is a critical point with

$$\nabla^2 h(\mathbf{U}) = \begin{bmatrix}4a + \frac{8}{a+2} - 6 & \frac{8}{a+2} - 6\\ \frac{8}{a+2} - 6 & 4a + \frac{8}{a+2} - 6\end{bmatrix},$$

which has eigenvalues

$$\lambda_1 = \frac{4(a-2)(a+1)}{a+2}\begin{cases}<0, & a\in[0,2),\\ >0, & a>2,\end{cases}$$

and $\lambda_2 = 4a > 0$. We conclude that this $\mathbf{U}$ is a strict saddle point when $a < 2$ and a spurious local minimum when $a > 2$. This weighted symmetric matrix factorization problem (5.9) satisfies the restricted strong convexity and smoothness condition (5.3) with constants $\alpha = \|\mathbf{\Omega}\|_{\min}^2 = 1$ and $\beta = \|\mathbf{\Omega}\|_{\max}^2 = 1 + a$ (where $\|\mathbf{\Omega}\|_{\min}$ and $\|\mathbf{\Omega}\|_{\max}$ represent the smallest and largest entries in $\mathbf{\Omega}$; see Section 5.3.3). Thus, we have a counter example which demonstrates the existence of spurious local minima when $\frac{\beta}{\alpha} > 3$.

Remark 5.3.4. We finally remark that although Theorem 5.3.1 requires the additional regularizer (5.4), empirical evidence (see experiments in Section 5.5) shows we can get rid of this regularizer for many iterative algorithms with random initialization.

We prove Theorem 5.3.1 in Section 5.4. Before proceeding, we present two stylized applications of Theorem 5.3.1 in matrix sensing and weighted low-rank approximation.

5.3.3 Stylized Applications

5.3.3.1 Matrix Sensing

We first consider the implication of Theorem 5.3.1 in the matrix sensing problem where

$$f(\mathbf{X}) = \frac{1}{2} \left\| \mathcal{A}\left(\mathbf{X} - \mathbf{X}^\star\right) \right\|_2^2.$$

Here $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ is a known measurement operator satisfying the following restricted isometry property.

Definition 5.3.1. *(Restricted Isometry Property (RIP) [65]) The map $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ satisfies the r-RIP with constant δ_r if*

$$(1 - \delta_r) \left\|\mathbf{X}\right\|_F^2 \leq \left\|\mathcal{A}(\mathbf{X})\right\|^2 \leq (1 + \delta_r) \left\|\mathbf{X}\right\|_F^2 \tag{5.10}$$

holds for any $n \times m$ matrix $\mathbf{X}$ with $\operatorname{rank}(\mathbf{X}) \leq r$.

Note that, in this case, the gradient of $f(\mathbf{X})$ at $\mathbf{X}^\star$ is

$$\nabla f(\mathbf{X}^\star) = \mathcal{A}^*\mathcal{A}(\mathbf{X}^\star - \mathbf{X}^\star) = \mathbf{0},$$

which implies that $\mathbf{X}^\star$ is a critical point of $f(\mathbf{X})$. The Hessian quadrature form $\nabla^2 f(\mathbf{X})[\mathbf{Y}, \mathbf{Y}]$ for any $n \times m$ matrices $\mathbf{X}$ and $\mathbf{Y}$ is given by

$$\nabla^2 f(\mathbf{X})[\mathbf{Y}, \mathbf{Y}] = \left\|\mathcal{A}(\mathbf{Y})\right\|^2.$$

If $\mathcal{A}$ satisfies the $4r$-restricted isometry property with constant δ_{4r}, then $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (5.3) with constants $\alpha = 1 - \delta_{4r}$ and $\beta = 1 - \delta_{4r}$ since

$$(1 - \delta_{4r}) \left\|\mathbf{Y}\right\|_F^2 \leq \left\|\mathcal{A}(\mathbf{Y})\right\|^2 \leq (1 + \delta_{4r}) \left\|\mathbf{Y}\right\|_F^2$$

for any rank-$4r$ matrix $\mathbf{Y}$. Now, applying Theorem 5.3.1, we can characterize the geometry for the following matrix sensing problem with the factorization approach:

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}, \mathbf{V}\in\mathbb{R}^{n\times r}}{\text{minimize}} \frac{1}{2} \left\|\mathcal{A}(\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star)\right\|_2^2 + g(\mathbf{U}, \mathbf{V}), \tag{5.11}$$

where $g(\mathbf{U}, \mathbf{V})$ is the added regularizer defined in (5.4).

Corollary 5.3.1. *Suppose $\mathcal{A}$ satisfies the $4r$-RIP with constant $\delta_{4r} \leq \frac{1}{5}$, and set $\mu \leq \frac{1-\delta_{4r}}{16}$. Then the objective function in* (5.11) *has no spurious local minima and satisfies the strict saddle property.*

This result follows directly from Theorem 5.3.1 by noting that $\frac{\beta}{\alpha} = \frac{1+\delta_{4r}}{1-\delta_{4r}} \leq 1.5$ if $\delta_{4r} \leq \frac{1}{5}$. We remark that Park et al. [106, Theorem 4.3] provided a similar geometric result for (5.11). Compared to their result which requires $\delta_{4r} \leq \frac{1}{100}$, our result has a much weaker requirement on the RIP of the measurement operator.

5.3.3.2 Weighted Low-Rank Matrix Factorization

We now consider the implication of Theorem 5.3.1 in the weighted matrix factorization problem [130], where

$$f(\mathbf{X}) := \frac{1}{2} \left\| \mathbf{\Omega} \circ (\mathbf{X} - \mathbf{X}^\star) \right\|_F^2 .$$

Here $\mathbf{\Omega}$ is an $n \times m$ weight matrix consisting of positive elements and $\circ$ denotes the point-wise product between two matrices. In this case, the gradient of $f(\mathbf{X})$ at $\mathbf{X}^\star$ is

$$\nabla f(\mathbf{X}^\star) = \mathbf{\Omega} \circ \mathbf{\Omega} \circ (\mathbf{X}^\star - \mathbf{X}^\star) = \mathbf{0},$$

which implies that $\mathbf{X}^\star$ is a critical point of $f(\mathbf{X})$. The Hessian quadrature form $\nabla^2 f(\mathbf{X})[\mathbf{Y}, \mathbf{Y}]$ for any $n \times m$ matrices $\mathbf{X}$ and $\mathbf{Y}$ is given by

$$\nabla^2 f(\mathbf{X})[\mathbf{Y}, \mathbf{Y}] = \left\| \mathbf{\Omega} \circ \mathbf{Y} \right\|_F^2 .$$

Thus $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (5.3) with constants $\alpha = \|\mathbf{\Omega}\|_{\min}^2$ and $\beta = \|\mathbf{\Omega}\|_{\max}^2$ since

$$\|\mathbf{\Omega}\|_{\min}^2 \|\mathbf{Y}\|_F^2 \leq \|\mathbf{\Omega} \circ \mathbf{Y}\|_F^2 \leq \|\mathbf{\Omega}\|_{\max}^2 \|\mathbf{Y}\|_F^2 ,$$

where $\|\mathbf{\Omega}\|_{\min}$ and $\|\mathbf{\Omega}\|_{\max}$ represent the smallest and largest entries in $\mathbf{\Omega}$, respectively. Now we consider the following weighted matrix factorization problem:

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{V} \in \mathbb{R}^{n \times r}}{\text{minimize}} \frac{1}{2} \left\| \mathbf{\Omega} \circ (\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star) \right\|_F^2 + g(\mathbf{U}, \mathbf{V}), \tag{5.12}$$

where $g(\mathbf{U}, \mathbf{V})$ is the added regularizer defined in (5.4). For an arbitrary weight matrix $\mathbf{\Omega}$, it is proven that the weighted low-rank factorization can be NP-hard [131] and has spurious local minima. When the elements in the weight matrix $\mathbf{\Omega}$ are concentrated, it is expected that (5.12) can be efficiently solved by a number of iterative optimization algorithms as it is close to an (unweighted) matrix factorization problem (where $\mathbf{\Omega}$ is a matrix of ones) which obeys the strict saddle property [104]. The following result characterizes the geometric structure in the objection function of (5.12) by directly applying Theorem 5.3.1.

Corollary 5.3.2. *Suppose* $\mathbf{\Omega}$ *satisfies* $\frac{\|\mathbf{\Omega}\|_{\max}^2}{\|\mathbf{\Omega}\|_{\min}^2} \leq 1.5$. *Set* $\mu \leq \frac{\|\mathbf{\Omega}\|_{\min}^2}{16}$. *Then the objective function in* (5.12) *has no spurious local minima and satisfies the strict saddle property.*

5.3.3.3 1-bit Matrix Completion

Finally, we consider the problem of completing a low-rank matrix from a subset of 1-bit measurements [132]. Given $\mathbf{X}^\diamond \in \mathbb{R}^{n \times m}$, a subset of indices $\Omega \subset [m] \times [n]$, and a differentiable function $q : \mathbb{R} \to [0, 1]$, we observe

$$Y_{i,j} = \begin{cases} +1 & \text{with probability } q(X_{i,j}^\diamond), \\ -1 & \text{with probability } 1 - q(X_{i,j}^\diamond), \end{cases} \tag{5.13}$$

for all $(i,j) \in \Omega$. Typical choices for q include the logistic regression model where $q(x) = \frac{e^x}{1+e^x}$ and the probit regression model where $q(x) = 1 - \Phi(-x/\sigma) = \Phi(x/\sigma)$. Here Φ is the cumulative distribution function (CDF) of a mean-zero Gaussian distribution with variance σ^2. In [132], the authors attempt to recover $\mathbf{X}^\diamond$ from the incomplete nonlinear measurements $\{Y_{ij}\}_{(i,j)\in\Omega}$ by minimizing the negative log-likelihood function

$$F_{\Omega,\mathbf{Y}}(\mathbf{X}) := -\sum_{(i,j)\in\Omega} \left(\mathbb{1}_{(Y_{i,j}=1)} \log(q(X_{i,j})) + \mathbb{1}_{(Y_{i,j}=-1)} \log(1-q(X_{i,j}))\right)$$

which results in a maximum likelihood (ML) estimate.

We note that $F_{\Omega,\mathbf{Y}}$ is a convex function for both the logistic model and the probit model. The following result also establishes that $F_{\Omega,\mathbf{Y}}$ satisfies the restricted strong convexity and smoothness condition if we observe full 1-bit measurements, i.e., $\Omega = [n] \times [m]$.

Lemma 5.3.1. *Suppose $\Omega = [n] \times [m]$. Let*

$$\alpha_{q,\gamma} = \min_{|x|\le\gamma} \min\left(\frac{(q'(x))^2 - q(x)q''(x)}{q^2(x)}, \frac{(q'(x))^2 + (1-q(x))q''(x)}{(1-q(x))^2}\right)$$

and

$$\beta_{q,\gamma} = \max_{|x|\le\gamma} \max\left(\frac{(q'(x))^2 - q(x)q''(x)}{q^2(x)}, \frac{(q'(x))^2 + (1-q(x))q''(x)}{(1-q(x))^2}\right).$$

Then $F_{\Omega,\mathbf{Y}}$ satisfies the restricted strong convexity and smoothness condition:

$$\alpha_{q,\gamma}\|\mathbf{G}\|_F^2 \le [\nabla^2 F_{\Omega,\mathbf{Y}}(\mathbf{X})](\mathbf{G},\mathbf{G}) \le \beta_{q,\gamma}\|\mathbf{G}\|_F^2$$

for any $\mathbf{G} \in \mathbb{R}^{n\times m}$ and $\|\mathbf{X}\|_\infty \le \gamma$.

The proof of Lemma 5.3.1 is given in Appendix D.1. Now we consider the logistic regression model where $q(x) = \frac{e^x}{1+e^x}$.

Corollary 5.3.3. *Suppose $\Omega = [n] \times [m]$ and $\gamma \le 1.3$. Consider the logistic regression model where $q(x) = \frac{e^x}{1+e^x}$. Then $F_{\Omega,\mathbf{Y}}$ satisfies the restricted strong convexity and smoothness condition with*

$$\frac{\beta_{q,\gamma}}{\alpha_{q,\gamma}} \le 1.5.$$

Proof of Corollary 5.3.3. Applying Lemma 5.3.1 with direct calculation gives

$$\alpha_{q,\gamma} = q'(\gamma) = \frac{e^\gamma}{(1+e^\gamma)^2},$$
$$\beta_{q,\gamma} = q'(0) = \frac{e^0}{(1+e^0)^2} = \frac{1}{4},$$

where $q'(x) = \frac{e^x}{(1+e^x)^2}$. Now if we restrict $\|\mathbf{X}\|_\infty \leq 1.3$, we have

$$\frac{\beta_{q,\gamma}}{\alpha_{q,\gamma}} = 4\frac{e^{1.3}}{(1+e^{1.3})^2} \leq 1.5.$$

□

Under the assumption that $\mathbf{X}^\diamond$ is low-rank, a nuclear norm constraint is utilized in [132] to force a low-rank solution. Corollary 5.3.3 implies that we can apply matrix factorization for 1-bit matrix recovery given that the elements of $\mathbf{X}$ are bounded. For the setting where Ω is only a subset of $[n] \times [m]$, [171] considered the 1-bit matrix *completion* problem with the rank constraint and established a stronger statistical recovery guarantee than that in [132]. Empirical evidence (see [171] and Section 5.5.3) supports that matrix factorization also works for 1-bit matrix completion.

5.4 Proof of Theorem 5.3.1

In this section, we provide a formal proof of Theorem 5.3.1. The main argument involves showing that each critical point of $\rho(\mathbf{W})$ either corresponds to the global solution of (5.1) or is a strict saddle whose Hessian $\nabla^2\rho(\mathbf{W})$ has a strictly negative eigenvalue. Specifically, we show that $\mathbf{W}$ is a strict saddle by arguing that the Hessian $\nabla^2\rho(\mathbf{W})$ has a strictly negative curvature along $\mathbf{\Delta} := \mathbf{W} - \mathbf{W}^\star\mathbf{R}$, i.e., $[\nabla^2\rho(\mathbf{W})](\mathbf{\Delta}, \mathbf{\Delta}) \leq -\tau\|\mathbf{\Delta}\|_F^2$ for some $\tau > 0$. Here $\mathbf{R}$ is an $r \times r$ orthonormal matrix such that the distance between $\mathbf{W}$ and $\mathbf{W}^\star$ rotated through $\mathbf{R}$ is as small as possible.

5.4.1 Supporting Results

We first present some useful results. The $(2r, 4r)$-restricted strong convexity and smoothness assumption (5.3) implies the following isometry property, whose proof is given in Appendix D.2.

Proposition 5.4.1. *Suppose the function $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (5.3) *with positive α and β. Then for any $n \times m$ matrices $\mathbf{Z}, \mathbf{G}, \mathbf{H}$ of rank at most $2r$, we have*

$$\left|\frac{2}{\alpha+\beta}[\nabla^2 f(\mathbf{Z})](\mathbf{G}, \mathbf{H}) - \langle \mathbf{G}, \mathbf{H}\rangle\right| \leq \frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{G}\|_F \|\mathbf{H}\|_F.$$

The following result provides an upper bound on the energy of the difference $\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}$ when projected onto the column space of $\mathbf{W}$. Its proof is given in Appendix D.3.

Lemma 5.4.1. *Suppose $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (5.3). *For any critical point $\mathbf{W}$ of* (5.5), *let $\mathbf{P}_\mathbf{W} \in \mathbb{R}^{(m+n)\times(m+n)}$ be the orthogonal projector onto the column space of $\mathbf{W}$. Then*

$$\left\|(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{P}_\mathbf{W}\right\|_F \leq 2\frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{X} - \mathbf{X}^\star\|_F.$$

We remark that Lemma 5.4.1 is a variant of [106, Lemma 3.2]. While the result there requires the $4r$-RIP condition of the objective function, our result depends on the $(2r, 4r)$-restricted strong convexity and smoothness condition. Our result is also slightly tighter than [106, Lemma 3.2].

In addition, for any matrices $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{n\times r}$, the following result relates the distance between $\mathbf{C}\mathbf{C}^\top$ and $\mathbf{D}\mathbf{D}^\top$ to the distance between $\mathbf{C}$ and $\mathbf{D}$.

Lemma 5.4.2. *For any matrices $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{n\times r}$ with ranks r_1 and r_2, respectively, let $\mathbf{R} = \arg\min_{\mathbf{R}'\in\mathcal{O}_r} \|\mathbf{C} - \mathbf{D}\mathbf{R}'\|_F$. Then*

$$\|\mathbf{C}\mathbf{C}^\top - \mathbf{D}\mathbf{D}^\top\|_F^2/\|\mathbf{C} - \mathbf{D}\mathbf{R}\|_F^2 \geq \max\left\{2(\sqrt{2}-1)\sigma_r^2(\mathbf{D}), \min\left\{\sigma_{r_1}^2(\mathbf{C}), \sigma_{r_2}^2(\mathbf{D})\right\}\right\}.$$

If $\mathbf{C} = \mathbf{0}$, then we have

$$\left\|\mathbf{C}\mathbf{C}^\top - \mathbf{D}\mathbf{D}^\top\right\|_F^2 \geq \sigma_{r_2}^2(\mathbf{D}) \|\mathbf{C} - \mathbf{D}\mathbf{R}\|_F^2.$$

We present one more useful result in the following Lemma.

Lemma 5.4.3. *[6, Lemma 5] For any matrices $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{n\times r}$, let $\mathbf{P}_\mathbf{C}$ be the orthogonal projector onto the range of $\mathbf{C}$. Let $\mathbf{R} = \arg\min_{\mathbf{R}'\in\mathcal{O}_r} \|\mathbf{C} - \mathbf{D}\mathbf{R}'\|_F$. Then*

$$\|\mathbf{C}\left(\mathbf{C} - \mathbf{D}\mathbf{R}\right)^\top\|_F^2 \leq \frac{1}{8}\|\mathbf{C}\mathbf{C}^\top - \mathbf{D}\mathbf{D}^\top\|_F^2 + (3 + \frac{1}{2(\sqrt{2}-1)})\|(\mathbf{C}\mathbf{C}^\top - \mathbf{D}\mathbf{D}^\top)\mathbf{P}_\mathbf{C}\|_F^2.$$

Finally, we provide the gradient and Hessian expressions for $\rho(\mathbf{W})$. The gradient of $\rho(\mathbf{W})$ is given by

$$\begin{aligned}\nabla_\mathbf{U}\rho(\mathbf{U},\mathbf{V}) &= \nabla f(\mathbf{X})\mathbf{V} + \mu\mathbf{U}(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}),\\ \nabla_\mathbf{V}\rho(\mathbf{U},\mathbf{V}) &= \nabla f(\mathbf{X})^\top\mathbf{U} - \mu\mathbf{V}(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}).\end{aligned}$$

Standard computations give the the Hessian quadrature form $[\nabla^2\rho(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta})$ for any $\mathbf{\Delta} = \begin{bmatrix}\mathbf{\Delta}_\mathbf{U}\\ \mathbf{\Delta}_\mathbf{V}\end{bmatrix}$ where $\mathbf{\Delta}_\mathbf{U} \in \mathbb{R}^{n\times r}, \mathbf{\Delta}_\mathbf{V} \in \mathbb{R}^{m\times r}$:

$$\begin{aligned}[\nabla^2\rho(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta}) = &[\nabla^2 f(\mathbf{X})](\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top, \mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top)\\ &+ 2\langle\nabla f(\mathbf{X}), \mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\rangle + [\nabla^2 g(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta}),\end{aligned}$$

where

$$[\nabla^2 g(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta}) = \mu\langle\widehat{\mathbf{W}}^\top\mathbf{W}, \widehat{\mathbf{\Delta}}^\top\mathbf{\Delta}\rangle + \mu\langle\widehat{\mathbf{W}}\widehat{\mathbf{\Delta}}^\top, \mathbf{\Delta}\mathbf{W}^\top\rangle + \mu\langle\widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \mathbf{\Delta}\mathbf{\Delta}^\top\rangle.$$

5.4.2 The Formal Proof

Proof of Theorem 5.3.1. Any critical point $\mathbf{W}$ of $\rho(\mathbf{W})$ satisfies $\nabla\rho(\mathbf{W}) = \mathbf{0}$, i.e.,

$$\nabla f(\mathbf{X})\mathbf{V} + \mu\mathbf{U}\left(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right) = \mathbf{0}, \tag{5.14}$$
$$\nabla f(\mathbf{X})^\top\mathbf{U} - \mu\mathbf{V}\left(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right) = \mathbf{0}. \tag{5.15}$$

By (5.15), we obtain

$$\mathbf{U}^\top\nabla f(\mathbf{X}) = \mu\left(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right)\mathbf{V}^\top.$$

Multiplying (5.14) by $\mathbf{U}^\top$ and plugging in the expression for $\mathbf{U}^\top\nabla f(\mathbf{X})$ from the above equation $\mathbf{V}^\top$ gives

$$(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V})\mathbf{V}^\top\mathbf{V} + \mathbf{U}^\top\mathbf{U}(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}) = \mathbf{0},$$

which further implies

$$\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}\mathbf{V}^\top\mathbf{V}.$$

Note that $\mathbf{U}^\top\mathbf{U}$ and $\mathbf{V}^\top\mathbf{V}$ are the principal square roots (i.e., PSD square roots) of $\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top\mathbf{U}$ and $\mathbf{V}^\top\mathbf{V}\mathbf{V}^\top\mathbf{V}$, respectively. Utilizing the result that a PSD matrix has a unique principal square root [172], we obtain

$$\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}. \tag{5.16}$$

Thus, we can simplify (5.14) and (5.15) by

$$\nabla_{\mathbf{U}}\rho(\mathbf{U},\mathbf{V}) = \nabla f(\mathbf{X})\mathbf{V} = \mathbf{0}, \tag{5.17}$$
$$\nabla_{\mathbf{V}}\rho(\mathbf{U},\mathbf{V}) = \nabla f(\mathbf{X})^\top\mathbf{U} = \mathbf{0}. \tag{5.18}$$

Now we turn to prove the strict saddle property and that there are no spurious local minima.

First, note that as guaranteed by Proposition 5.3.1, $\mathbf{X}^\star$ is the unique $n \times m$ matrix with rank at most r. Also the gradient of $f(\mathbf{X})$ vanishes at $\mathbf{X}^\star$ since (5.1) is an unconstraint optimization problem. Denote the set of critical points of $\rho(\mathbf{W})$ by

$$\mathcal{C} := \left\{\mathbf{W} \in \mathbb{R}^{(n+m)\times r} : \nabla\rho(\mathbf{W}) = \mathbf{0}\right\}.$$

We separate $\mathcal{C}$ into two subsets:

$$\mathcal{C}_1 := \mathcal{C} \cap \left\{\mathbf{W} \in \mathbb{R}^{(n+m)\times r} : \mathbf{U}\mathbf{V}^\top = \mathbf{X}^\star\right\},$$
$$\mathcal{C}_2 := \mathcal{C} \cap \left\{\mathbf{W} \in \mathbb{R}^{(n+m)\times r} : \mathbf{U}\mathbf{V}^\top \neq \mathbf{X}^\star\right\},$$

satisfying $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$. Since any critical point $\mathbf{W}$ satisfies (5.16), $g(\mathbf{W})$ achieves its global minimum at $\mathbf{W}$. Also $f(\mathbf{X})$ achieves its global minimum at $\mathbf{X}^\star$. We conclude that $\mathbf{W}$ is the globally optimal solution of ρ for any $\mathbf{W} \in \mathcal{C}_1$. If we show that any $\mathbf{W} \in \mathcal{C}_2$ is a strict saddle, then we prove that there are no spurious local minima as well as the strict saddle property. Thus, the remaining part is to show that $\mathcal{C}_2$ is the set of strict saddles.

To show that $\mathcal{C}_2$ is the set of strict saddles, it is sufficient to find a direction $\mathbf{\Delta}$ along which the Hessian has a strictly negative curvature for each of these points. We construct $\mathbf{\Delta} = \mathbf{W} - \mathbf{W}^\star\mathbf{R}$, the difference from $\mathbf{W}$ to its nearest global factor $\mathbf{W}^\star$, where

$$\mathbf{R} = \underset{\mathbf{R}' \in \mathcal{O}_r}{\arg\min} \left\| \mathbf{W} - \mathbf{W}^\star\mathbf{R}' \right\|_F .$$

Such $\mathbf{\Delta}$ satisfies $\mathbf{\Delta} \neq \mathbf{0}$ since $\mathbf{X} \neq \mathbf{X}^\star$ implying $\mathbf{W}\mathbf{W}^\top \neq \mathbf{W}^\star\mathbf{W}^{\star\top}$. Then we evaluate the Hessian bilinear form along the direction $\mathbf{\Delta}$:

$$\begin{aligned}
[\nabla^2 \rho(\mathbf{W})](\mathbf{\Delta}, \mathbf{\Delta}) &= 2 \underbrace{\langle \nabla f(\mathbf{X}), \mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top \rangle}_{\Pi_1} \\
&+ \underbrace{[\nabla^2 f(\mathbf{X})](\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top, \mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top)}_{\Pi_2} + \mu \underbrace{\langle \widehat{\mathbf{W}}\widehat{\mathbf{\Delta}}^\top, \mathbf{\Delta}\mathbf{W}^\top \rangle}_{\Pi_3} + \mu \underbrace{\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \mathbf{\Delta}\mathbf{\Delta}^\top \rangle}_{\Pi_4} .
\end{aligned} \tag{5.19}$$

The following result (which is proved in Appendix D.5) states that Π_1 is strictly negative, while the remaining terms are relatively small, though they may be nonnegative:

$$\begin{aligned}
&\Pi_1 \leq -\alpha \left\| \mathbf{X} - \mathbf{X}^\star \right\|_F^2 , \quad \Pi_2 \leq \beta \|\mathbf{W}\mathbf{\Delta}^\top\|_F^2, \\
&\Pi_3 \leq \|\mathbf{W}\mathbf{\Delta}^\top\|_F^2, \quad \Pi_4 \leq 2 \left\| \mathbf{X} - \mathbf{X}^\star \right\|_F^2 .
\end{aligned} \tag{5.20}$$

Now, substituting (5.20) into (5.19) gives

$$\begin{aligned}
[\nabla^2 \rho(\mathbf{W})](\mathbf{\Delta}, \mathbf{\Delta}) &= 2\Pi_1 + \Pi_2 + \mu\Pi_3 + \mu\Pi_4 \\
&\leq -2\alpha \|\mathbf{X} - \mathbf{X}^\star\|_F^2 + (\beta + \mu) \cdot \|\mathbf{W}\mathbf{\Delta}^\top\|_F^2 + 2\mu \left\| \mathbf{X} - \mathbf{X}^\star \right\|_F^2 \\
&\overset{(i)}{\leq} (-2\alpha + 2\mu) \left\| \mathbf{X} - \mathbf{X}^\star \right\|_F^2 + (\beta + \mu)(\frac{1}{2} + (12 + \frac{2}{\sqrt{2} - 1})(\frac{\beta - \alpha}{\beta + \alpha})^2) \left\| \mathbf{X} - \mathbf{X}^\star \right\|_F^2 \\
&\overset{(ii)}{\leq} -0.2\alpha \left\| \mathbf{X} - \mathbf{X}^\star \right\|_F^2 ,
\end{aligned} \tag{5.21}$$

where (i) utilizes Lemmas 5.4.1 and 5.4.3, (ii) utilizes the following inequality (which is proved in Appendix D.6)

$$\left\| \mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^\star \right\|_F^2 \leq 4 \left\| \mathbf{X} - \mathbf{X}^\star \right\|_F^2 , \tag{5.22}$$

and (ii) holds because $\frac{\beta}{\alpha} \leq 1.5$ and $\mu \leq \frac{1}{16}\alpha$. Thus, if $\mathbf{X} \neq \mathbf{X}^\star$, $\left[\nabla^2 \rho(\mathbf{X})\right](\mathbf{\Delta}, \mathbf{\Delta})$ is always negative. This implies that $\mathbf{W}$ is a strict saddle.

To complete the proof, we utilize Lemma 5.4.2 to further bound the last term in (5.21):

$$
\begin{aligned}
[\nabla^2\rho(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta}) &\leq -0.05\alpha\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\|_F^2 \\
&\leq -0.05\alpha\|\mathbf{\Delta}\|_F^2 \begin{cases} 2(\sqrt{2}-1)\sigma_r^2(\mathbf{W}^\star), & r = r^\star, \\ \min\left\{\sigma_{r^c}^2(\mathbf{W}), \sigma_{r^\star}^2(\mathbf{W}^\star)\right\}, & r > r^\star, \\ \sigma_{r^\star}^2(\mathbf{W}^\star), & r_c = 0, \end{cases}
\end{aligned}
$$

where r^c is the rank of $\mathbf{W}$, the fist inequality utilizes (5.22), and the second inequality follows from Lemma 5.4.2. We complete the proof of Theorem 5.3.1 by noting that $\sigma_\ell^2(\mathbf{W}^\star) = 2\sigma_\ell(\mathbf{X}^\star)$ for all $\ell \in \{1,\ldots,r^\star\}$ since

$$
\mathbf{W}^\star = \begin{bmatrix} \mathbf{Q}_{\mathbf{U}^\star}\mathbf{\Sigma}^{\star 1/2} \\ \mathbf{Q}_{\mathbf{V}^\star}\mathbf{\Sigma}^{\star 1/2} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_{\mathbf{U}^\star}/\sqrt{2} \\ \mathbf{Q}_{\mathbf{V}^\star}/\sqrt{2} \end{bmatrix} \left(\sqrt{2}\mathbf{\Sigma}^{\star 1/2}\right)\mathbf{I}
$$

is an SVD of $\mathbf{W}^\star$, where we recall that $\mathbf{X}^\star = \mathbf{Q}_{\mathbf{U}^\star}\mathbf{\Sigma}^\star\mathbf{Q}_{\mathbf{V}^\star}^\top$ is an SVD of $\mathbf{X}^\star$. □

Remark 5.4.1. From (5.21), we observe that a smaller μ yields a more negative bound on $\left[\nabla^2\rho(\mathbf{X})\right](\mathbf{\Delta},\mathbf{\Delta})$. This can be explained intuitively as follows. First note that any critical point $\mathbf{W}$ satisfies (5.16) provided $\mu > 0$, no matter how large or small μ is. The Hessian information about $g(\mathbf{W})$ is represented by the terms Π_3 and Π_4. We have

$$
\begin{aligned}
\Pi_3 + \Pi_4 &= \left\langle \widehat{\mathbf{W}}\widehat{\mathbf{\Delta}}^\top, \mathbf{\Delta}\mathbf{W}^\top \right\rangle + \left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \mathbf{\Delta}\mathbf{\Delta}^\top \right\rangle \\
&= \left\langle \widehat{\mathbf{W}}^\top\mathbf{\Delta}, \mathbf{\Delta}^\top\widehat{\mathbf{W}} \right\rangle + \left\langle \widehat{\mathbf{W}}^\top\mathbf{\Delta}, \widehat{\mathbf{W}}^\top\mathbf{\Delta} \right\rangle \\
&= \left\langle \widehat{\mathbf{W}}^\top\mathbf{\Delta}, \widehat{\mathbf{W}}^\top\mathbf{\Delta} + \mathbf{\Delta}^\top\widehat{\mathbf{W}} \right\rangle \\
&\geq 0,
\end{aligned}
$$

where the last line holds since for any $r \times r$ matrix $\mathbf{A}$,

$$
\begin{aligned}
\left\langle \mathbf{A}, \mathbf{A} + \mathbf{A}^\top \right\rangle &= \frac{1}{2}\left\langle \mathbf{A} + \mathbf{A}^\top, \mathbf{A} + \mathbf{A}^\top \right\rangle + \frac{1}{2}\left\langle \mathbf{A} - \mathbf{A}^\top, \mathbf{A} + \mathbf{A}^\top \right\rangle \\
&= \frac{1}{2}\left\|\mathbf{A} + \mathbf{A}^\top\right\|_F^2 \geq 0.
\end{aligned}
$$

Thus the Hessian of ρ evaluated at any critical point $\mathbf{W}$ is a PSD matrix[12] instead of having a negative eigenvalue. In low-rank, PSD matrix optimization problems, the corresponding objective function (without any regularizer such as $g(\mathbf{W})$) is proved to have the strict saddle property [6, 125]. Therefore, $h(\mathbf{W})$ is also expected to have the strict saddle property, and so is $\rho(\mathbf{W})$ when μ is small, i.e., the Hessian of $g(\mathbf{W})$ has little influence on the Hessian of $\rho(\mathbf{W})$ when μ is small. Our results also indicate that when the restricted strict convexity constant α is not provided a priori, we can always choose a small μ to ensure the strict saddle property of $\rho(\mathbf{W})$ is met, and hence we are guaranteed the global convergence of a number of local search algorithms applied to (5.5).

[12] This can also be observed since any critical point $\mathbf{W}$ is a global minimum point of $\rho(\mathbf{W})$, which directly indicates that $\nabla^2\rho(\mathbf{W}) \succeq \mathbf{0}$.

5.5 Experiments

In this section, we present a set of experiments on matrix sensing, matrix completion, and 1-bit matrix completion to demonstrate the performance of iterative algorithms for low-rank matrix optimization. Unless noted otherwise, we denote the matrix factorization approach by NVX and use the minFunc package[13] to perform the local search algorithms for the factored problem.

5.5.1 Matrix Sensing

We first present some experiments to illustrate the performance of local search algorithms for the matrix sensing problem with the factorization approach (5.11). In these experiments, we set $n = 50$, $m = 50$ and vary the rank r from 1 to 19. We generate a rank-r $n \times m$ random matrix $\mathbf{X}^\star$ by setting $\mathbf{X}^\star = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top$ where $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ are respectively $n \times r$ and $m \times r$ matrixes of normally distributed random numbers. We then obtain p random measurements $\mathbf{y} = \mathcal{A}(\mathbf{X}^\star)$ with

$$y_i = \langle \mathbf{X}^\star, \mathbf{Y}_i \rangle ,$$

where the entries of each $n \times m$ matrix $\mathbf{Y}_i$ are independent and identically distributed (i.i.d.) normal random variables with zero mean and variance $\frac{1}{p}$ for $i \in \{1, 2, \dots p\}$. For each pair of r and the number of measurements, 10 Monte Carlo trials are carried out and for each trial, and we claim matrix recovery to be successful if the relative reconstruction error satisfies

$$\frac{\|\mathbf{X}^\star - \widehat{\mathbf{X}}\|_F}{\|\mathbf{X}^\star\|_F} \leq 10^{-4},$$

where we denote by $\widehat{\mathbf{X}}$ the reconstructed matrix. Figure 5.1 displays the phase transition for factorized gradient descent starting from a random initialization, the singular value projection (SVP) method proposed in [173] which requires a SVD in each iteration, and the convex approach which solves

$$\begin{aligned} &\underset{\mathbf{X}}{\text{minimize}} \ \|\mathbf{X}\|_* \\ &\text{subject to } \ \mathbf{y} = \mathcal{A}(\mathbf{X}). \end{aligned} \tag{5.23}$$

We see that there are only negligible differences between the different approaches for matrix sensing; these approaches also have very similar performance guarantees when the Gaussian sensing operator $\mathcal{A}$ satisfies the RIP [156]. We note that with or without the regularizer g as defined in (5.4), local search algorithms have similar performance with random initialization. Hence, throughout all of the experiments, we simply discard the regularizer g, but we stress that identical performance is observed if we have this regularizer g.

[13] Software available at
https://www.cs.ubc.ca/∼schmidtm/Software/minFunc.html

The previous experiments suppose that r is known for SVP and the matrix factorization approach. We note, however, that our result in Theorem 5.3.1 also covers the over-parameterization case where $r > r^\star$. To illustrate the possible influence of over-parameterization, we generate a rank-$r^\star$ random matrix $\mathbf{X}^\star \in \mathbb{R}^{n\times m}$ with $r^\star = 4$ and $n = m = 50$ and obtain $p = 4Rn$ random measurements (so that the measurement operator $\mathcal{A}$ satisfies the RIP of rank R), where $R = 7$. We then solve the matrix factorization problem[14] with $r = 4, 5, 6, 7$ and display the corresponding convergence results in Figure 5.2. As can been seen, the matrix factorization approach converges to the target matrix $\mathbf{X}^\star$ in both the exact-parameterization and over-parameterization cases. However, we also observe that it converges slower in the over-parameterization case (i.e., $r > r^\star$) than in the exact-parameterization case (i.e., $r = r^\star$).

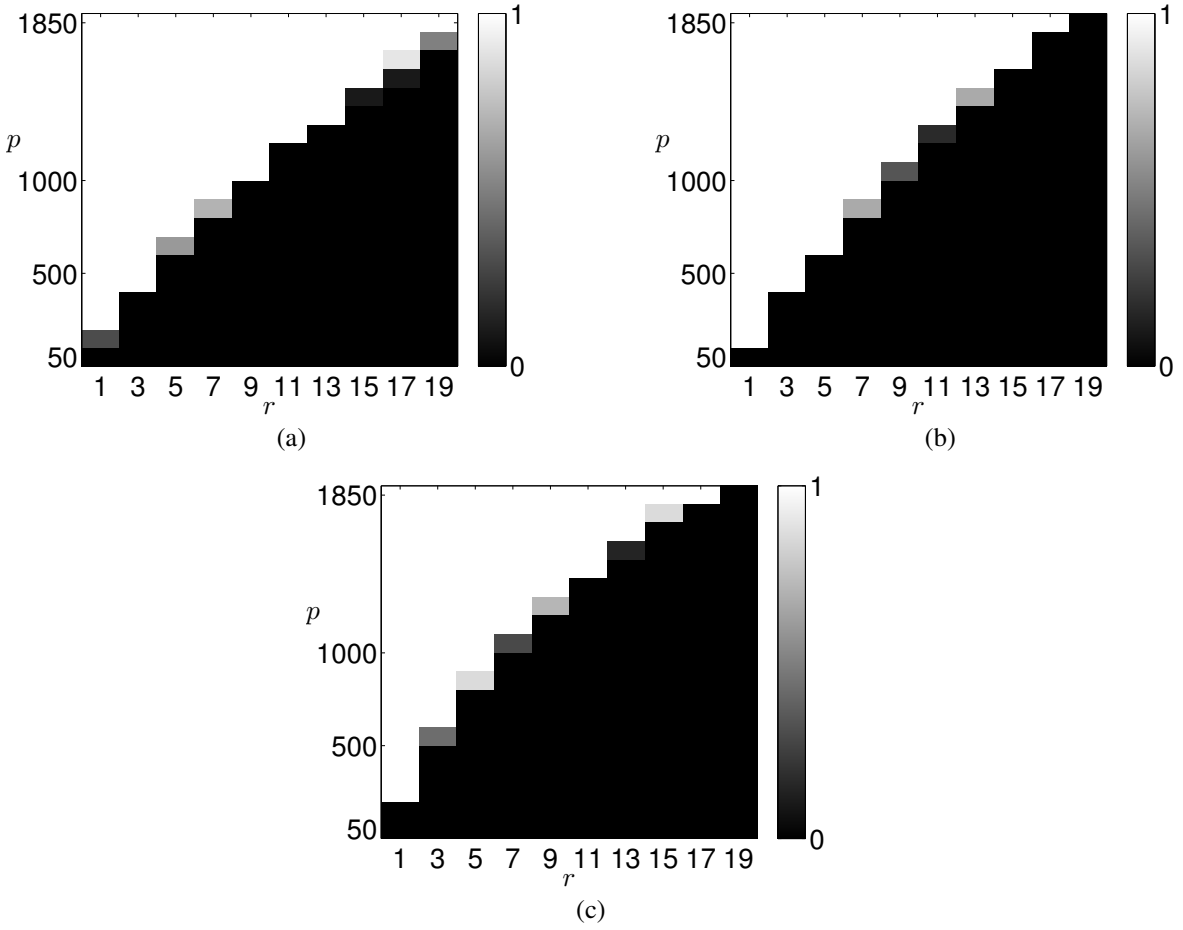


Figure 5.1: Rate of success for matrix sensing by (a) solving the factorized problem (5.11) with gradient descent; (b) SVP [173]; (c) solving the convex problem (5.23).

[14]To avoid tuning the parameters (such as step-size) for different r, we use the minFunc package with the default setting, which solves the factored problem by the "LBFGS" algorithm [174].

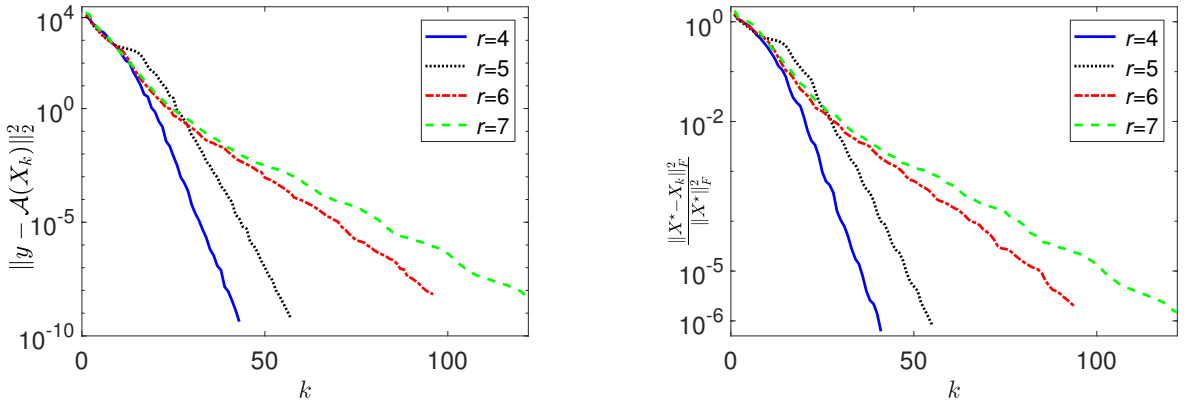


Figure 5.2: The performance in terms of (a) objective value and (b) the relative Frobenius norm of the error versus the iteration k for the matrix factorization approach solving matrix sensing with $r^\star = 4, n = m = 50, p = 4Rn, R = 7$ and r varying from $r^\star$ to R.

5.5.2 Matrix Completion

We compare the performance of the matrix factorization approach with SVP [173], the convex approach, and singular value thresholding[15] (SVT) [175] for matrix completion where we want to recover a low-rank matrix $\mathbf{X}^\star$ from incomplete measurements $\{X^\star_{ij}\}_{(i,j)\in\Omega}$, where $\Omega \subset [n] \times [m]$. Let $\mathcal{P}_\Omega$ denote the projection onto the index set Ω. The convex approach (denoted by CVX) attempts to use the nuclear norm as a convex relaxation of the rankness and solves

$$\begin{aligned} &\underset{\mathbf{X}}{\text{minimize}} \ \|\mathbf{X}\|_* \\ &\text{subject to } \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^\star). \end{aligned} \tag{5.24}$$

To make the recovery of $\mathbf{X}^\star$ well-posed, we require $\mathbf{X}^\star$ to be incoherent such that the information in $\mathbf{X}$ is not concentrated in a small number of entries [153]. A matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ with singular value decomposition $\mathbf{X} = \mathbf{L}\mathbf{\Sigma}\mathbf{Q}^\top$ is u-incoherent if [173, Definition 2.1]

$$\max_{ij} |L_{ij}| \le \sqrt{\frac{u}{n}}, \ \max_{ij} |Q_{ij}| \le \sqrt{\frac{u}{m}}.$$

Though $\mathcal{P}_\Omega$ does not satisfy the r-RIP (5.10) for all low-rank matrices $\mathbf{X}$, it satisfies the RIP when restricted to low-rank incoherent matrices.

Theorem 5.5.1. *[173, Theorem 4.2] Without loss of generality, assume $n \ge m$. There exists a constant $C \ge 0$ such that for $\Omega \in [n] \times [m]$ chosen according to the Bernouli model with density greater than $Cu^2r^2 \log n/\delta^2 m$, with probability at least $1 - e^{-n\log n}$, the RIP holds for all μ-incoherent matrices $\mathbf{X}$ of rank at most r.*

[15] Software available at http://svt.stanford.edu/

Thus, if local search algorithms (such as gradient descent) start with a random initialization and the iterates remain incoherent, then Theorem 5.3.1 guarantees the global convergence of the matrix factorization approach with these algorithms. We note that this hypothesis is also required for SVP [173]. Though we can add a regularizer for incoherence as in [100], empirical evidence supports this hypothesis that the iterates in gradient descent are incoherent.

In the first set of experiments, we set $n = m = 100$ and vary the rank r from 1 to 30. Similar to the setup for matrix sensing in Section 5.5.1, we generate a rank-r random matrix and randomly obtain p entries, i.e., $|\Omega| = p$. Figure 5.3 displays the phase transition for gradient descent with a random initialization, SVP [173], singular value thresholding (SVT) [175], and the convex approach. As can been seen, the matrix factorization approach has similar phase transition to SVP, and is slightly better than SVT and the convex approach in terms of the number of measurements needed for successful recovery.

In the second set of experiments, we set $r = 5$ and $p = 3r(2n - r)$ (3 times the number of degrees of freedom within a rank-r $n \times n$ matrix), and vary n from 40 to 5120. We compare the time needed for the four approaches in Figure 5.4; our matrix factorization approach is much faster than the other methods. The time savings for the matrix factorization approach comes from avoiding performing the SVD, which is needed both for SVT and SVP in each iteration. We also observe that convex approach has the highest computational complexity and is not scalable (which is the reason that we only present its time for n up to 640).

5.5.3 1-bit Matrix Completion

In the last set of experiments, we compare the performance of the matrix factorization approach with the convex approach[16] in [132] for 1-bit matrix completion. We first note that to make the recovery problem well-posed, a constraint on $\|\mathbf{X}\|_\infty$ (the entry-wise maximum of the matrix $\mathbf{X}$) is applied in [132] to require that the matrix is not too "spiky". Instead of using the constraint on $\|\mathbf{X}\|_\infty$, we add a smooth regularizer $\|\mathbf{X}\|_F^2$ and turn to minimize the following objective function

$$f_{\Omega,\mathbf{Y}}(\mathbf{X}) = F_{\Omega,\mathbf{Y}}(\mathbf{X}) + \frac{\eta}{2}\|\mathbf{X}\|_F^2,$$

which is also a convex function over $\mathbf{X}$ and satisfies a similar restricted strong convexity and smoothness condition to $F_{\Omega,\mathbf{Y}}$ in Lemma 5.3.1. In the case where we only observe part of the entries, then in light of Theorem 5.5.1, the corresponding objective function is expected to satisfy the strong convexity and smoothness condition for all incoherent matrices. Thus, we factorize $\mathbf{X}$ into $\mathbf{U}\mathbf{V}^\top$ and solve the following optimization problem over the $n \times r$ and $m \times r$ matrices $\mathbf{U}$ and $\mathbf{V}$:

$$\underset{\mathbf{U},\mathbf{V}}{\text{minimize}}\, \rho_{\Omega,\mathbf{Y}}(\mathbf{U},\mathbf{V}) = f_{\Omega,\mathbf{Y}}(\mathbf{U}\mathbf{V}^\top). \tag{5.25}$$

[16] Software available at http://mdav.ece.gatech.edu/software/

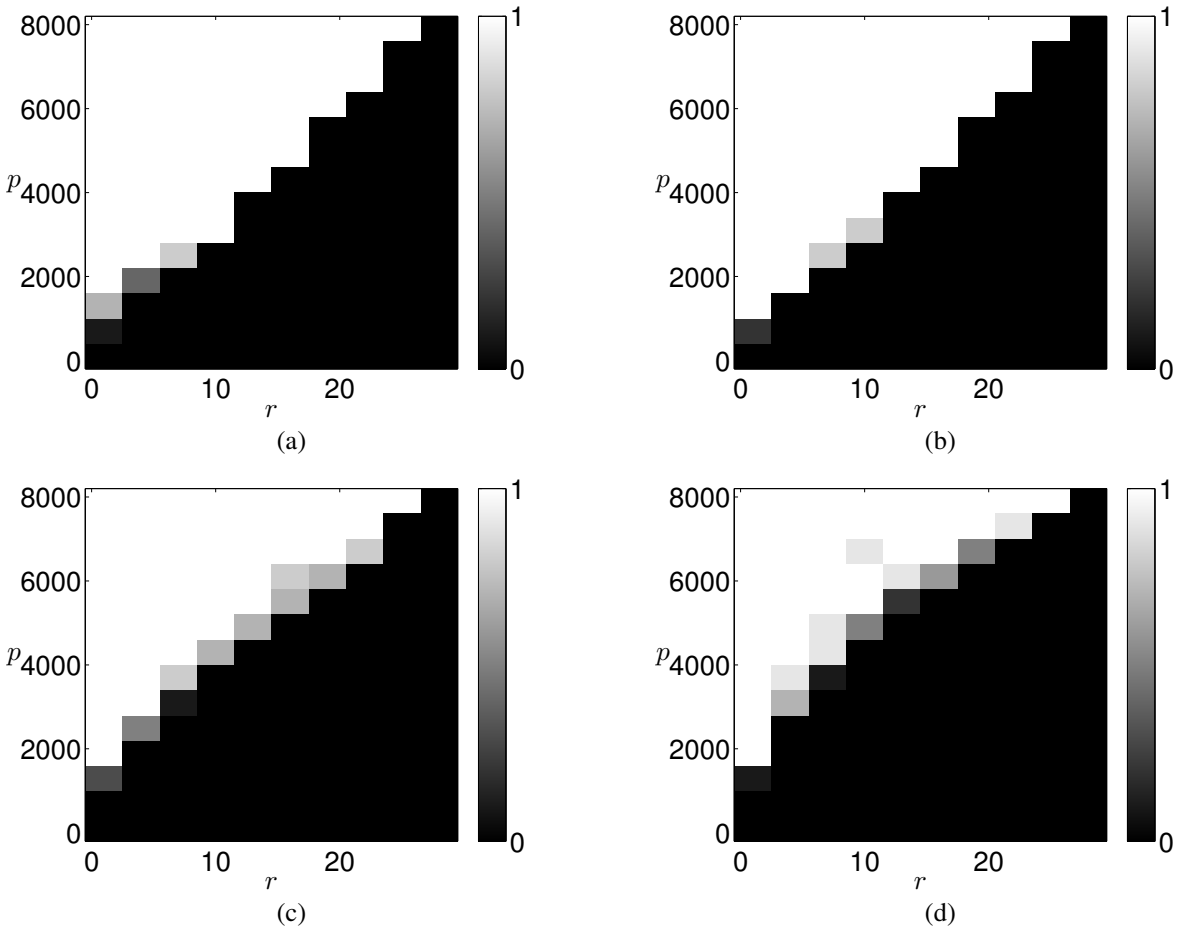


Figure 5.3: Rate of success for matrix sensing by (a) the matrix factorization approach with gradient descent; (b) SVP [173]; (c) solving the convex problem (5.24); (d) SVT [173].

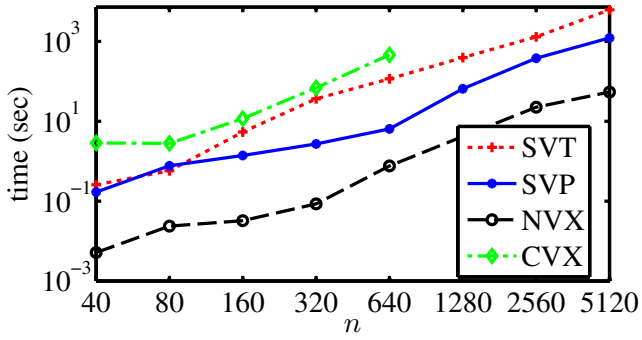


Figure 5.4: Average computation time needed for different algorithms solving matrix completion.

To evaluate the performance of this factorization approach on 1-bit matrix completion, we generate $n \times r$ matrices $\mathbf{U}^\diamond$ and $\mathbf{V}^\diamond$ with entries drawn i.i.d. from a uniform distribution on $[-\frac{1}{2}, \frac{1}{2}]$ and construct a random $n \times n$ matrix $\mathbf{X}^\diamond$ with rank r. Similar to the setup in [132], the matrix is then scaled so that $\|\mathbf{X}^\diamond\| = 1$. We obtain 1-bit observations

$\{Y_{i,j}\}_{(i,j)\in\Omega}$ by adding Gaussian noise of variance σ^2 and recording the sign of the resulting value (5.13), where the subset of indices Ω is chosen at random with $\mathrm{E}\,|\Omega| = p$. We compare the performance of the factorization approach and the convex approach [132] over a range of different values of n, p, r or σ. Figure 5.5(a)-(d) show the normalized squared Frobenius norm of the error $\frac{\|\widehat{\mathbf{X}}-\mathbf{X}^\diamond\|}{\|\mathbf{X}^\diamond\|_F^2}$ (where $\widehat{\mathbf{X}}$ denotes the reconstructed matrix) and average the results over 10 draws of Monte Carlo trials. We observe that matrix factorization approach has slightly better performance than the convex approach for 1-bit matrix completion [132]. Note that this phenomenon (the factorization approach having better performance) is also observed in [171]. We repeat these experiments but obtaining 1-bit observations with the logistic regression model where $g(x) = \frac{e^x}{1+e^x}$ for (5.13) and display the results in Figure 5.6.

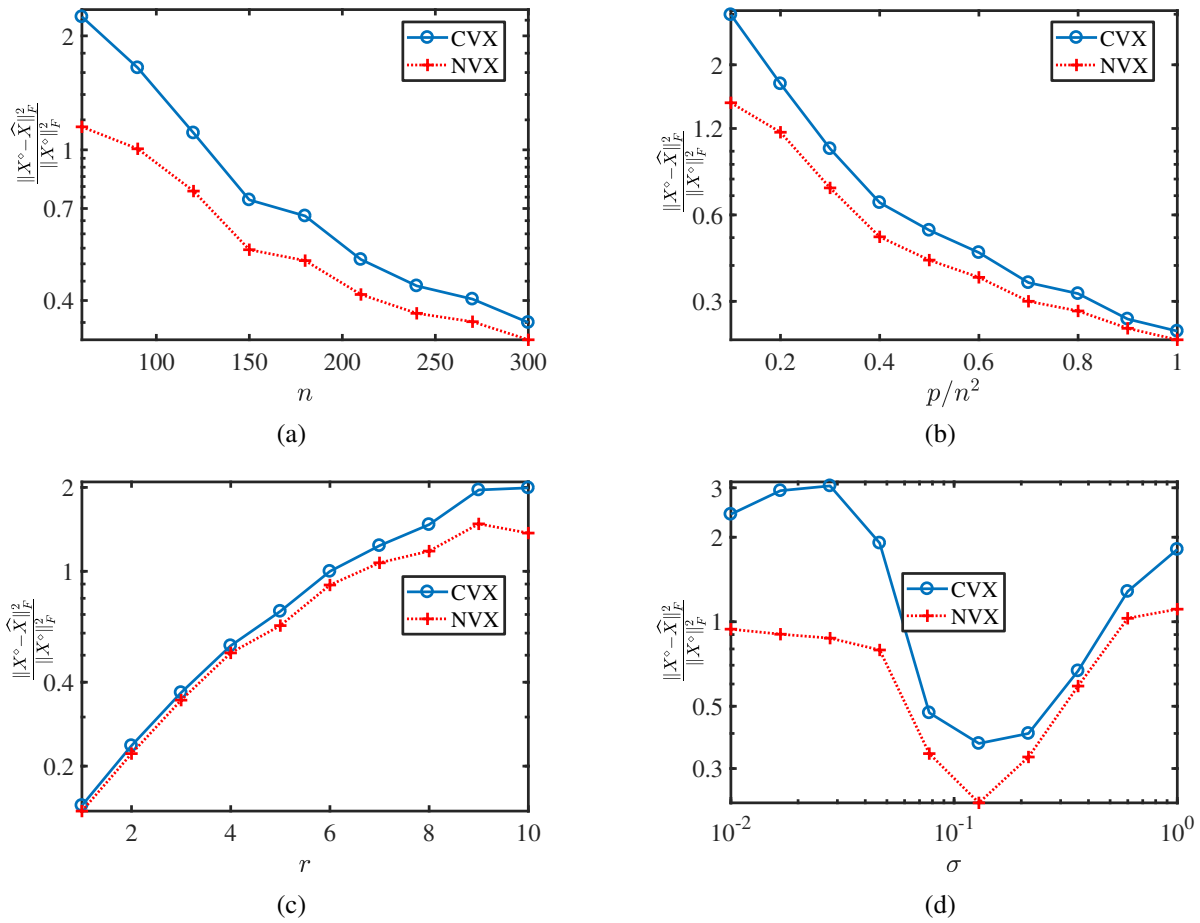


Figure 5.5: The performance in terms of the relative Frobenius norm of the error for the matrix factorization approach (denoted by NVX) and the convex approach in [132] (denoted by CVX) for solving the 1-bit matrix completion with probit regression model and (a) varying n and $\sigma = 0.3$, $r = 7$, $p = 0.5n^2$; (b) varying p and $\sigma = 0.3$, $n = 200$, $r = 7$; (c) varying r and $\sigma = 0.3$, $n = 200$, $p = 0.25n^2$; (d) varying σ and $n = 200$, $r = 4$, $p = 0.25n^2$. The results are plotted in the log scale.

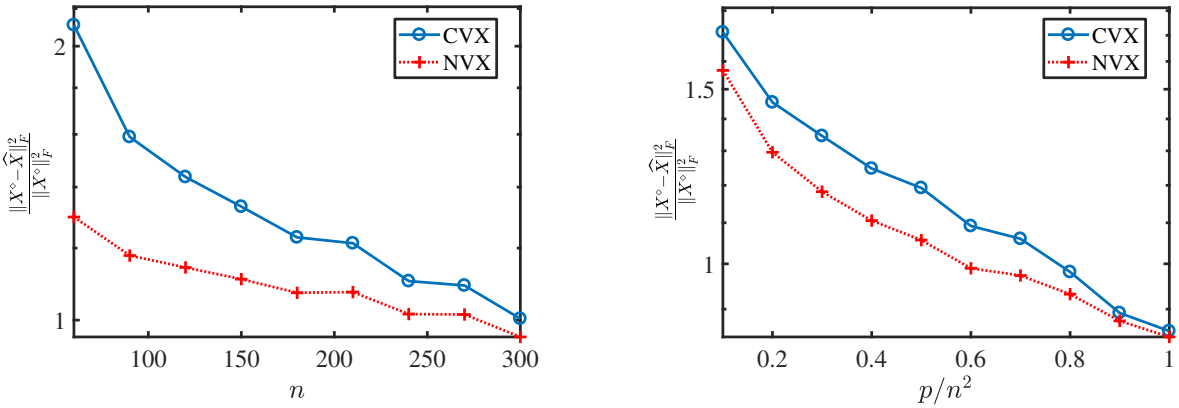


Figure 5.6: The performance in terms of the relative Frobenius norm of the error for the matrix factorization approach (denoted by NVX) and the convex approach in [132] (denoted by CVX) for solving the 1-bit matrix completion with logistic regression model and (a) varying n and $r = 2$, $p = 0.5n^2$; (b) varying p and $n = 200$, $r = 2$. The results are plotted in the log scale.

5.6 Conclusion

This work considers low-rank matrix optimization on general (nonsymmetric and rectangular) matrices with general objective functions. By focusing on general objective functions, we provide a unifying framework for low-rank matrix optimizations with the factorization approach. Although the resulting optimization problem is not convex, we show that the reformulated objection function has a simple landscape: there are no spurious local minima and any critical point not being a local minimum is a strict saddle such that the Hessian evaluated at this point has a strictly negative eigenvalue. These properties guarantee that a number of iterative optimization algorithms (such as gradient descent and the trust region method) will converge to the global optimum from a random initialization.

CHAPTER 6

THE GLOBAL OPTIMIZATION GEOMETRY OF LOW-RANK MATRIX OPTIMIZATION

This work considers general rank-constrained optimization problems that minimize a general objective function $f(\mathbf{X})$ over the set of rectangular $n \times m$ matrices that have rank at most r. To tackle the rank constraint and also to reduce the computational burden, we factorize $\mathbf{X}$ into $\mathbf{U}\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ are $n \times r$ and $m \times r$ matrices, respectively, and then optimize over the small matrices $\mathbf{U}$ and $\mathbf{V}$. We characterize the global optimization geometry of the nonconvex factored problem and show that the corresponding objective function satisfies the robust strict saddle property as long as the original objective function f satisfies restricted strong convexity and smoothness properties, ensuring global convergence of many local search algorithms (such as noisy gradient descent) in polynomial time for solving the factored problem. We also provide a comprehensive analysis for the optimization geometry of a matrix factorization problem where we aim to find $n \times r$ and $m \times r$ matrices $\mathbf{U}$ and $\mathbf{V}$ such that $\mathbf{U}\mathbf{V}^\top$ approximates a given matrix $\mathbf{X}^\star$. Aside from the robust strict saddle property, we show that the objective function of the matrix factorization problem has no spurious local minima and obeys the strict saddle property not only for the exact-parameterization case where $\mathrm{rank}(\mathbf{X}^\star) = r$, but also for the over-parameterization case where $\mathrm{rank}(\mathbf{X}^\star) < r$ and the under-parameterization case where $\mathrm{rank}(\mathbf{X}^\star) > r$. These geometric properties imply that a number of iterative optimization algorithms (such as gradient descent) converge to a global solution with random initialization.

6.1 Introduction

Low-rank matrices arise in a wide variety of applications throughout science and engineering, ranging from quantum tomography [149], signal processing [176], machine learning [151, 169], and so on; see [124] for a comprehensive review. In all of these settings, we often encounter the following rank-constrained optimization problem:

$$\begin{aligned} &\underset{\mathbf{X} \in \mathbb{R}^{n \times m}}{\text{minimize}}\ f(\mathbf{X}), \\ &\text{subject to } \mathrm{rank}(\mathbf{X}) \le r, \end{aligned} \tag{6.1}$$

where the objective function $f : \mathbb{R}^{n \times m} \to \mathbb{R}$ is smooth.

Whether the objective function f is convex or nonconvex, the rank constraint renders low-rank matrix optimizations of the form (6.1) highly nonconvex and computationally NP-hard in general [155]. Significant efforts have been devoted to transforming (6.1) into a convex problem by replacing the rank constraint with one involving the nuclear norm. This strategy has been widely utilized in matrix inverse problems [65] arising in signal processing [124], machine learning [157], and control [155]. With convex analysis techniques, nuclear norm minimization has been proved to provide optimal performance in recovering low-rank matrices [177]. However, in spite of the optimal performance,

solving nuclear norm minimization is very computationally expensive even with specialized first-order algorithms. For example, the singular value thresholding algorithm [175] requires performing an expensive singular value decomposition (SVD) in each iteration, making it computationally prohibitive in large-scale settings. This prevents nuclear norm minimization from scaling to practical problems.

To relieve the computational bottleneck, recent studies propose to factorize the variable into $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, and optimize over the $n \times r$ and $m \times r$ matrices $\mathbf{U}$ and $\mathbf{V}$ rather than the $n \times m$ matrix $\mathbf{X}$. The rank constraint in (6.1) then is automatically satisfied through the factorization. This strategy is usually referred to as the Burer-Monteiro type decomposition after the authors in [119, 159]. Plugging this parameterization of $\mathbf{X}$ in (6.1), we can recast the program into the following one:

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \; h(\mathbf{U},\mathbf{V}) := f(\mathbf{U}\mathbf{V}^\top). \tag{6.2}$$

The bilinear nature of the parameterization renders the objective function of (6.2) nonconvex. Hence, it can potentially have spurious local minima (i.e., local minimizers that are not global minimizers) or even saddle points. With technical innovations in analyzing the landscape of nonconvex functions, however, several recent works have shown that the factored objective function $h(\mathbf{U},\mathbf{V})$ in certain matrix inverse problems has no spurious local minima [100, 106, 125].

6.1.1 Summary of Results and Outline

In this work, we provide a comprehensive geometric analysis for solving general low-rank optimizations of the form (6.1) using the factorization approach (6.2). Our work actually rests on the recent works [97, 114, 126, 178, 179] ensuring a number of iterative optimization methods (such as gradient descent) converge to a local minimum with random initialization provided the problem satisfies the so-called strict saddle property (see Definition 6.2.3 in Section 6.2). If the objective function further obeys the robust strict saddle property [114] (see Definition 6.2.4 in Section 6.2) or belongs to the class of so-called $\mathcal{X}$ functions [97], the recent works [97, 114] show that many local search algorithms can converge to a local minimum in polynomial time. The implications of this line of work have had a tremendous impact on a number of nonconvex problems in applied mathematics, signal processing, and machine learning.

We begin this chapter in Section 6.2 with the notions of strict saddle, strict saddle property, and robust strict saddle property. Considering that many invariant functions are not strongly convex (or even convex) in any neighborhood around a local minimum point, we then provide a revised robust strict saddle property[17] requiring a regularity condition (see Definition 6.2.8 in Section 6.2) rather than strong convexity near the local minimum points (which is one of the

[17] A similar notion of a revised robust strict saddle property has also been utilized in [126], which shows that noisy gradient descent converges to a local minimum in a number iterations that depends only poly-logarithmically on the dimension. In a nutshell, [126] has a different focus than this work: the focus in [126] is on providing convergence analysis of a noisy gradient descent algorithm with a robust strict saddle property, while in the present work, we establish a robust strict saddle property for the nonsymmetric matrix factorization and more general low-rank optimization (including matrix sensing) problems with the factorization approach.

requirements for the strict saddle property). The stochastic gradient descent algorithm is guaranteed to converge to a local minimum point in polynomial time for problems satisfying the revised robust strict saddle property [114, 126].

In Section 6.3, we consider the geometric analysis for solving general low-rank optimizations of the form (6.1) using the factorization approach (6.2). Provided the objective function f satisfies certain restricted strong convexity and smoothness conditions, we show that the low-rank optimization problem with the factorization (6.2) (with an additional regularizer—see Section 6.3 for the details) obeys the revised robust strict saddle property. In Section 6.3.3, we consider a stylized application in matrix sensing where the measurement operator satisfies the restricted isometry property (RIP) [65]. In the case of Gaussian measurements, as guaranteed by this robust strict saddle property, a number of iterative optimizations can find the unknown matrix $\mathbf{X}^\star$ of rank r in polynomial time with high probability when the number of measurements exceeds a constant times $(n+m)r^2$.

Our main approach for analyzing the optimization geometry of (6.2) is based on the geometric analysis for the following non-square low-rank matrix factorization problem: given $\mathbf{X}^\star \in \mathbb{R}^{n\times m}$,

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}^{m\times r}}{\text{minimize}} \left\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\right\|_F^2. \tag{6.3}$$

In particular, we show the optimization geometry for the low-rank matrix factorization problem (6.3) is preserved for the general low-rank optimization (6.2) under certain restricted strong convexity and smoothness conditions on f. Thus, in Appendix E.1, we provide a comprehensive geometric analysis for (6.3), which can be viewed as an important foundation of many popular matrix factorization problems such as the matrix sensing problem and matrix completion. We show that the low-rank matrix factorization problem (6.3) (with an additional regularizer) has no spurious local minima and obeys the strict saddle property—that is the objective function in (6.3) has a directional negative curvature at all critical points but local minima—not only for the exact-parameterization case where $\mathrm{rank}(\mathbf{X}^\star) = r$, but also for the over-parameterization case where $\mathrm{rank}(\mathbf{X}^\star) < r$ and the under-parameterization case where $\mathrm{rank}(\mathbf{X}^\star) > r$. The strict saddle property and lack of spurious local minima ensure that a number of local search algorithms applied to the matrix factorization problem (6.3) converge to global optima which correspond to the best rank-r approximation to $\mathbf{X}^\star$. Further, we completely analyze the low-rank matrix factorization problem (6.3) for the exact-parameterization case and show that it obeys the revised robust strict saddle property.

6.1.2 Relation to Existing Work

Unlike the objective functions of convex optimizations that have simple landscapes, such as where all local minimizers are global ones, the objective functions of general nonconvex programs have much more complicated landscapes. In recent years, by exploiting the underlying optimization geometry, a surge of progress has been made in providing theoretical justifications for matrix factorization problems such as (6.2) using a number of previously heuristic algorithms (such as alternating minimization, gradient descent, and the trust region method). Typical ex-

amples include phase retrieval [98, 180, 181], blind deconvolution [182, 183], dictionary learning [184, 185], phase synchronization [186] and matrix sensing and completion [100–102, 109, 140, 166, 187].

These iterative algorithms can be sorted into two categories based on whether a good initialization is required. One set of algorithms consist of two steps: initialization and local refinement. Provided the function satisfies a regularity condition or similar properties, a good guess lying in the attraction basin of the global optimum can lead to global convergence of the following iterative step. We can obtain such initializations by spectral methods for phase retrieval [180], phase synchronization [186] and low-rank matrix recovery problems [102, 103, 108, 109]. As we have mentioned, a regularity condition is also adopted in the revised robust strict saddle property.

Another category of works attempt to analyze the landscape of the objective functions in a larger space rather than the regions near the global optima. We can further separate these approaches into two types based on whether they involve the strict saddle property or the robust strict saddle property. The strict saddle property and lack of spurious local minima are proved for low-rank, positive semidefinite (PSD) matrix recovery [125] and completion [100], PSD matrix optimization problems with generic objective functions [6], low-rank non-square matrix estimation from linear observations [106], low-rank nonsquare optimization problems with generic objective functions [8] and generic nuclear norm reggularized problems [92]. The strict saddle property along with the lack of spurious local minima ensures a number of iterative algorithms such as gradient descent [114] and the trust region method [170] converge to the global minimum with random initialization [114, 178, 185].

A few other works which are closely related to our work attempt to study the *global geometry* by characterizing the landscapes of the objective functions in the whole space rather than the regions near the global optima or all the critical points. As we discussed before, a number of local search algorithms are guaranteed to find a local optimum (which is also the global optimum if there are no spurious local minima) because of this robust strict saddle property. In [114], the authors proved that tensor decompostion problems satisfy this robust strict saddle property. Sun et al. [98] studied the global geometry of the phase retrieval problem. The very recent work in [104] analyzed the global geometry for PSD low-rank matrix factorization of the form (6.3) and the related matrix sensing problem when the rank is exactly parameterized (i.e., $r = \mathrm{rank}(\mathbf{X}^\star)$). The factorization approach for matrix inverse problems with quadratic loss functions is considered in [101]. We extend this line by considering general rank-constrained optimization problems including a set of matrix inverse problems.

Finally, we remark that our work is also closely related to the recent works in low-rank matrix factorization of the form (6.3) and its variants [8, 100–102, 104, 106, 109, 125, 140]. As we discussed before, most of these works except [101, 104] (but including [106] which also focuses on nonsymmetric matrix sensing) only characterize the geometry either near the global optima or all the critical points. Instead, we characterize the *global* geometry for general (rather than PSD) low-rank matrix factorization and sensing. Because the analysis is different, the proof strategy in the present work is also very different than that of [8, 106]. The results for PSD matrix sensing in [104]

build heavily on the concentration properties of Gaussian measurements, while our results for matrix sensing depend on the RIP of the measurement operator and thus can be applied to other matrix sensing problems whose measurement operator is not necessarily from a Gaussian measurement ensemble. Also, [101] considers matrix inverse problems with quadratic loss functions and its proof strategy is very different than that in the present work: the proof in [101] is specified to quadratic loss functions, while we consider the rank-constrained optimization problem with general objective functions in (6.1) and our proof utilizes the fact that the gradient and Hessian of the low-rank matrix sensing are respectively very close to those in low-rank matrix factorization. Furthermore, in terms of the matrix factorization, we show that the objective function in (6.3) obeys the strict saddle property and has no spurious local minima not only for exact-parameterization ($r = \mathrm{rank}(\mathbf{X}^\star)$), but also for over-parameterization ($r > \mathrm{rank}(\mathbf{X}^\star)$) and under-parameterization ($r < \mathrm{rank}(\mathbf{X}^\star)$). Local (rather than global) geometry results for exact-parameterization and under-parameterization are also covered in [8]. As noted above, the work in [101, 104] for low-rank matrix factorization only focuses on exact-parameterization ($r = \mathrm{rank}(\mathbf{X}^\star)$). The under-parameterization implies that we can find the best rank-r approximation to $\mathbf{X}^\star$ by many efficient iterative optimization algorithms such as gradient descent.

6.1.3 Notation

Before proceeding, we first briefly introduce some notation used throughout the chapter. The symbols $\mathbf{I}$ and $\mathbf{0}$ respectively represent the identity and zero matrices with appropriate sizes. Also $\mathbf{I}_n$ is used to denote the $n \times n$ identity matrix. The set of $r \times r$ orthonormal matrices is denoted by $\mathcal{O}_r := \{\mathbf{R} \in \mathbb{R}^{r\times r} : \mathbf{R}^\top\mathbf{R} = \mathbf{I}\}$. For any natural number n, we let $[n]$ or $1:n$ denote the set $\{1, 2, ..., n\}$. We use $|\Omega|$ denote the cardinality (i.e., the number of elements) of a set Ω. MATLAB notations are adopted for matrix indexing; that is, for the $n \times m$ matrix $\mathbf{A}$, its (i, j)-th element is denoted by $\mathbf{A}[i, j]$, its i-th row (or column) is denoted by $\mathbf{A}[i, :]$ (or $\mathbf{A}[:, i]$), and $\mathbf{A}[\Omega_1, \Omega_2]$ refers to a $|\Omega_1| \times |\Omega_2|$ submatrix obtained by taking the elements in rows Ω_1 of columns Ω_2 of matrix $\mathbf{A}$. Here $\Omega_1 \subset [n]$ and $\Omega_2 \subset [n]$. We use $a \gtrsim b$ (or $a \lesssim b$) to represent that there is a constant so that $a \geq \text{Const} \cdot b$ (or $a \leq \text{Const} \cdot b$).

If a function $h(\mathbf{U}, \mathbf{V})$ has two arguments, $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{V} \in \mathbb{R}^{m\times r}$, we occasionally use the notation $h(\mathbf{W})$ when we put these two arguments into a new one as $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$. For a scalar function $f(\mathbf{Z})$ with a matrix variable $\mathbf{Z} \in \mathbb{R}^{n\times m}$, its gradient is an $n\times m$ matrix whose (i, j)-th entry is $[\nabla f(\mathbf{Z})][i, j] = \frac{\partial f(\mathbf{Z})}{\partial \mathbf{Z}[i,j]}$ for all $i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, m\}$. The Hessian of $f(\mathbf{Z})$ can be viewed as an $nm \times nm$ matrix $[\nabla^2 f(\mathbf{Z})][i, j] = \frac{\partial^2 f(\mathbf{Z})}{\partial \mathbf{z}[i]\partial \mathbf{z}[j]}$ for all $i, j \in \{1, \ldots, nm\}$, where $\mathbf{z}[i]$ is the i-th entry of the vectorization of $\mathbf{Z}$. An alternative way to represent the Hessian is by a bilinear form defined via $[\nabla^2 f(\mathbf{Z})](\mathbf{A}, \mathbf{B}) = \sum_{i,j,k,l} \frac{\partial^2 f(\mathbf{Z})}{\partial \mathbf{Z}[i,j]\partial \mathbf{Z}[k,\ell]} \mathbf{A}[i, j]\mathbf{B}[k, \ell]$ for any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times m}$. These two notations will be used interchangeably whenever the specific form can be inferred from context.

6.2 Preliminaries

In this section, we provide a number of important definitions in optimization and group theory. To begin, suppose $h(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable.

Definition 6.2.1 (Critical points)**.** *A point* $\mathbf{x}$ *is a critical point of* $h(\mathbf{x})$ *if* $\nabla h(\mathbf{x}) = \mathbf{0}$.

Definition 6.2.2 (Strict saddles; or ridable saddles in [185])**.** *A critical point* $\mathbf{x}$ *is a strict saddle if the Hessian matrix evaluated at this point has a strictly negative eigenvalue, i.e.,* $\lambda_{\min}(\nabla^2 h(\mathbf{x})) < 0$.

Definition 6.2.3 (Strict saddle property [114])**.** *A twice differentiable function satisfies the strict saddle property if each critical point either corresponds to a local minimum or is a strict saddle.*

Intuitively, the strict saddle property requires a function to have a directional negative curvature at all of the critical points but local minima. This property allows a number of iterative algorithms such as noisy gradient descent [114] and the trust region method [170] to further decrease the function value at all the strict saddles and thus converge to a local minimum.

In [114], the authors proposed a noisy gradient descent algorithm for the optimization of functions satisfying the robust strict saddle property.

Definition 6.2.4 (Robust strict saddle property [114])**.** *Given* $\alpha, \gamma, \epsilon, \delta$*, a twice differentiable* $h(\mathbf{x})$ *satisfies the* $(\alpha, \gamma, \epsilon, \delta)$*-robust strict saddle property if for every point* $\mathbf{x}$ *at least one of the following applies:*

1. *There exists a local minimum point* $\mathbf{x}^\star$ *such that* $\|\mathbf{x}^\star - \mathbf{x}\| \leq \delta$*, and the function* $h(\mathbf{x}')$ *restricted to a* 2δ *neighborhood of* $\mathbf{x}^\star$ *(i.e.,* $\|\mathbf{x}^\star - \mathbf{x}'\| \leq 2\delta$*) is* α*-strongly convex;*
2. $\lambda_{\min}\left(\nabla^2 h(\mathbf{x})\right) \leq -\gamma$*;*
3. $\|\nabla h(\mathbf{x})\| \geq \epsilon$.

In words, the above robust strict saddle property says that for any point whose gradient is small, then either the Hessian matrix evaluated at this point has a strictly negative eigenvalue, or it is close to a local minimum point. Thus the robust strict saddle property not only requires that the function obeys the strict saddle property, but also that it is well-behaved (i.e., strongly convex) near the local minima and has large gradient at the points far way to the critical points.

Intuitively, when the gradient is large, the function value will decrease in one step by gradient descent; when the point is close to a saddle point, the noise introduced in the noisy gradient descent could help the algorithm escape the saddle point and the function value will also decrease; when the point is close to a local minimum point, the algorithm then converges to a local minimum. Ge et al. [114] rigorously showed that the noisy gradient descent algorithm

(see [114, Algorithm 1]) outputs a local minimum in a polynomial number of steps if the function $h(\mathbf{x})$ satisfies the robust strict saddle property.

It is proved in [114] that tensor decompostion problems satisfy this robust strict saddle property. However, requiring the local strong convexity prohibits the potential extension of the analysis in [114] for the noisy gradient descent algorithm to many other problems, for which it is not possible to be strongly convex in any neighborhood around the local minimum points. Typical examples include the matrix factorization problems due to the rotational degrees of freedom for any critical point. This motivates us to weaken the local strong convexity assumption relying on the approach used by [102, 180] and to provide the following revised robust strict saddle property for such problems. To that end, we list some necessary definitions related to groups and invariance of a function under the group action.

Definition 6.2.5 (Definition 7.1 [188])**.** *A (closed) binary operation, $\circ$, is a law of composition that produces an element of a set from two elements of the same set. More precisely, let $\mathcal{G}$ be a set and $a_1, a_2 \in \mathcal{G}$ be arbitrary elements. Then $(a_1, a_2) \to a_1 \circ a_2 \in \mathcal{G}$.*

Definition 6.2.6 (Definition 7.2 [188])**.** *A* group *is a set $\mathcal{G}$ together with a (closed) binary operation $\circ$ such that for any elements $a, a_1, a_2, a_3 \in \mathcal{G}$ the following properties hold:*

- *Associative property: $a_1 \circ (a_2 \circ a_3) = (a_1 \circ a_2) \circ a_3$.*
- *There exists an identity element $e \in \mathcal{G}$ such that $e \circ a = a \circ e = a$.*
- *There is an element $a^{-1} \in \mathcal{G}$ such that $a^{-1} \circ a = a \circ a^{-1} = e$.*

With this definition, it is common to denote a group just by $\mathcal{G}$ without saying the binary operation $\circ$ when it is clear from the context.

Definition 6.2.7. *Given a function $h(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ and a group $\mathcal{G}$ of operators on $\mathbb{R}^n$, we say h is invariant under the group action (or under an element a of the group) if*

$$h(a(\mathbf{x})) = h(\mathbf{x})$$

for all $\mathbf{x} \in \mathbb{R}^n$ and $a \in \mathcal{G}$.

Suppose the group action also preserves the energy of $\mathbf{x}$, i.e., $\|a(\mathbf{x})\| = \|\mathbf{x}\|$ for all $a \in \mathcal{G}$. Since for any $\mathbf{x} \in \mathbb{R}^n$, $h(a(\mathbf{x})) = h(\mathbf{x})$ for all $a \in \mathcal{G}$, it is straightforward to stratify the domain of $h(\mathbf{x})$ into equivalent classes. The vectors in each of these equivalent classes differ by a group action. One implication is that when considering the distance of two points $\mathbf{x}_1$ and $\mathbf{x}_2$, it would be helpful to use the distance between their corresponding classes:

$$\begin{aligned} \operatorname{dist}(\mathbf{x}_1, \mathbf{x}_2) :&= \min_{a_1 \in \mathcal{G}, a_2 \in \mathcal{G}} \|a_1(\mathbf{x}_1) - a_2(\mathbf{x}_2)\| \\ &= \min_{a \in \mathcal{G}} \|\mathbf{x}_1 - a(\mathbf{x}_2)\|, \end{aligned} \tag{6.4}$$

where the second equality follows because $\|a_1(\mathbf{x}_1) - a_2(\mathbf{x}_2)\| = \|a_1(\mathbf{x}_1 - a_1^{-1} \circ a_2(\mathbf{x}_2))\| = \|\mathbf{x}_1 - a_1^{-1} \circ a_2(\mathbf{x}_2)\|$ and $a_1^{-1} \circ a_2 \in \mathcal{G}$. Another implication is that the function $h(\mathbf{x})$ cannot possibly be strongly convex (or even convex) in any neighborhood around its local minimum points because of the existence of the equivalent classes. Before presenting the revised robust strict saddle property for invariant functions, we list two examples to illuminate these concepts.

Example 1. As one example, consider the phase retrieval problem of recovering an n-dimensional complex vector $\mathbf{x}^\star$ from $\left\{y_i = \left|\mathbf{b}_i^{\mathrm{H}}\mathbf{x}^\star\right|, i = 1, \ldots, p\right\}$, the magnitude of its projection onto a collection of known complex vectors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p$ [98, 180]. The unknown $\mathbf{x}^\star$ can be estimated by solving the following natural least-squares formulation [98, 180]

$$\underset{\mathbf{x} \in \mathbb{C}^n}{\text{minimize}}\, h(\mathbf{x}) = \frac{1}{2p} \sum_{i=1}^{p} \left(y_i^2 - \left|\mathbf{b}_i^{\mathrm{H}}\mathbf{x}\right|^2\right)^2,$$

where we note that here the domain of $\mathbf{x}$ is $\mathbb{C}^n$. For this case, we denote the corresponding

$$\mathcal{G} = \{e^{j\theta} : \theta \in [0, 1)\}$$

and the group action as $a(\mathbf{x}) = e^{j\theta}\mathbf{x}$, where $a = e^{j\theta}$ is an element in $\mathcal{G}$. It is clear that $h(a(\mathbf{x})) = h(\mathbf{x})$ for all $a \in \mathcal{G}$. Due to this invariance of $h(\mathbf{x})$, it is impossible to recover the global phase factor of the unknown $\mathbf{x}^\star$ and the function $h(\mathbf{x})$ is not strongly convex in any neighborhood of $\mathbf{x}^\star$.

Example 2. As another example, we revisit the general factored low-rank optimization problem (6.2):

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{V} \in \mathbb{R}^{m \times r}}{\text{minimize}}\, h(\mathbf{U}, \mathbf{V}) = f(\mathbf{U}\mathbf{V}^\top).$$

We recast the two variables $\mathbf{U}, \mathbf{V}$ into $\mathbf{W}$ as $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$. For this example, we denote the corresponding $\mathcal{G} = \mathcal{O}_r$ and the group action on $\mathbf{W}$ as $a(\mathbf{W}) = \begin{bmatrix} \mathbf{U}\mathbf{R} \\ \mathbf{V}\mathbf{R} \end{bmatrix}$ where $a = \mathbf{R} \in \mathcal{G}$. We have that $h(a(\mathbf{W})) = h(\mathbf{W})$ for all $a \in \mathcal{G}$ since $\mathbf{U}\mathbf{R}(\mathbf{V}\mathbf{R})^\top = \mathbf{U}\mathbf{V}^\top$ for any $\mathbf{R} \in \mathcal{O}_r$. Because of this invariance, in general $h(\mathbf{W})$ is not strongly convex in any neighborhood around its local minimum points even though $f(\mathbf{X})$ is a strongly convex function; see [104] for the symmetric low-rank factorization problem and Theorem E.1.1 in Appendix E.1 for the nonsymmetric low-rank factorization problem.

In the examples illustrated above, due to the invariance, the function is not strongly convex (or even convex) in any neighborhood around its local minimum point and thus it is prohibitive to apply the standard approach in optimization to show the convergence in a small neighborhood around the local minimum point. To overcome this issue, Candès et al. [180] utilized the so-called regularity condition as a sufficient condition for local convergence of gradient descent

applied for the phase retrieval problem. This approach has also been applied for the matrix sensing problem [102] and semi-definite optimization [103].

Definition 6.2.8 (Regularity condition [102, 180]). *Suppose $h(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ is invariant under the group action of the given group $\mathcal{G}$. Let $\mathbf{x}^\star \in \mathbb{R}^n$ be a local minimum point of $h(\mathbf{x})$. Define the set $B(\delta, \mathbf{x}^\star)$ as*

$$B(\delta, \mathbf{x}^\star) := \{\mathbf{x} \in \mathbb{R}^n : \operatorname{dist}(\mathbf{x}, \mathbf{x}^\star) \leq \delta\},$$

where the distance $\operatorname{dist}(\mathbf{x}, \mathbf{x}^\star)$ is defined in (6.4)*. Then we say the function $h(\mathbf{x})$ satisfies the (α, β, δ)-regularity condition if for all $\mathbf{x} \in B(\delta, \mathbf{x}^\star)$, we have*

$$\langle \nabla h(\mathbf{x}), \mathbf{x} - a(\mathbf{x}^\star) \rangle \geq \alpha \operatorname{dist}(\mathbf{x}, \mathbf{x}^\star)^2 + \beta \|\nabla h(\mathbf{x})\|^2, \tag{6.5}$$

where $a = \arg\min_{a' \in \mathcal{G}} \|\mathbf{x} - a'(\mathbf{x}^\star)\|$.

We remark that (α, β) in the regularity condition (6.2.8) must satisfy $\alpha\beta \leq \frac{1}{4}$ since by applying Cauchy-Schwarz

$$\langle \nabla h(\mathbf{x}), \mathbf{x} - a(\mathbf{x}^\star) \rangle \leq \|\nabla h(\mathbf{x})\| \operatorname{dist}(\mathbf{x}, \mathbf{x}^\star)$$

and the inequality of arithmetic and geometric means

$$\alpha \operatorname{dist}^2(\mathbf{x}, \mathbf{x}^\star) + \beta \|\nabla h(\mathbf{x})\|^2 \geq 2\sqrt{\alpha\beta} \operatorname{dist}(\mathbf{x}, \mathbf{x}^\star) \|\nabla h(\mathbf{x})\|^2.$$

Lemma 6.2.1. *[102, 180] If the function $h(\mathbf{x})$ restricted to a δ neighborhood of $\mathbf{x}^\star$ satisfies the (α, β, δ)-regularity condition, then as long as gradient descent starts from a point $\mathbf{x}_0 \in B(\delta, \mathbf{x}^\star)$, the gradient descent update*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \nu \nabla h(\mathbf{x}_t)$$

with step size $0 < \nu \leq 2\beta$ obeys $\mathbf{x}_t \in B(\delta, \mathbf{x}^\star)$ and

$$\operatorname{dist}^2(\mathbf{x}_t, \mathbf{x}^\star) \leq (1 - 2\nu\alpha)^t \operatorname{dist}^2(\mathbf{x}_0, \mathbf{x}^\star)$$

for all $t \geq 0$.

The proof is given in [180]. To keep the chapter self-contained, we also provide the proof of Lemma 6.2.1 in Appendix E.2. We remark that the decreasing rate $1 - 2\nu\alpha \in [0, 1)$ since we choose $\nu \leq 2\beta$ and $\alpha\beta \leq \frac{1}{4}$.

Now we establish the following revised robust strict saddle property for invariant functions by replacing the strong convexity condition in Definition 6.2.4 with the regularity condition.

Definition 6.2.9 (Revised robust strict saddle property for invariant functions). *Given a twice differentiable $h(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ and a group $\mathcal{G}$, suppose $h(\mathbf{x})$ is invariant under the group action and the energy of $\mathbf{x}$ is also preserved under the group action, i.e., $h(a(\mathbf{x})) = h(\mathbf{x})$ and $\|a(\mathbf{x})\|_2 = \|\mathbf{x}\|_2$ for all $a \in \mathcal{G}$. Given $\alpha, \beta, \gamma, \epsilon, \delta$, $h(\mathbf{x})$ satisfies the $(\alpha, \beta, \gamma, \epsilon, \delta)$-robust strict saddle property if for any point $\mathbf{x}$ at least one of the following applies:*

1. *There exists a local minimum point* $\mathbf{x}^\star$ *such that* $\mathrm{dist}(\mathbf{x}, \mathbf{x}^\star) \leq \delta$*, and the function* $h(\mathbf{x}')$ *restricted to* 2δ *a neighborhood of* $\mathbf{x}^\star$ *(i.e.,* $\mathrm{dist}(\mathbf{x}', \mathbf{x}^\star) \leq 2\delta$*) satisfies the* $(\alpha, \beta, 2\delta)$*-regularity condition defined in Definition 6.2.8;*

2. $\lambda_{\min}\left(\nabla^2 h(\mathbf{x})\right) \leq -\gamma$*;*

3. $\|\nabla h(\mathbf{x})\| \geq \epsilon$*.*

Compared with Definition 6.2.4, the revised robust strict saddle property requires the local descent condition instead of strict convexity in a small neighborhood around any local minimum point. With the convergence guarantee in Lemma 6.2.1, the convergence analysis of the stochastic gradient descent algorithm in [114] for the robust strict saddle functions can also be applied for the revised robust strict saddle functions defined in Definition 6.2.9 with the same convergence rate.[18] We omit the details here and refer the reader to [126] for more details on this. In the rest of the chapter, the robust strict saddle property refers to the one in Definition 6.2.9.

6.3 Low-rank Matrix Optimization with the factorization approach

In this section, we consider the minimization of general rank-constrained optimization problems of the form (6.1) using the factorization approach (6.2) (which we repeat as follows):

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}, \mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \; h(\mathbf{U}, \mathbf{V}) = f(\mathbf{U}\mathbf{V}^\top),$$

where the rank constraint in (6.1) is automatically satisfied by the factorization approach. With necessary assumptions on f in Section 6.3.1, we provide geometric analysis of the factored problem in Section 6.3.2. We then present a stylized application in matrix sensing in Section 6.3.3.

6.3.1 Assumptions And Regularizer

Before presenting our main results, we lay out the necessary assumptions on the objective function $f(\mathbf{X})$. As is known, without any assumptions on the problem, even minimizing traditional quadratic objective functions is challenging. For this reason, we focus on problems satisfying the following two assumptions.

Assumption 6.3.1. $f(\mathbf{X})$ *has a critical point* $\mathbf{X}^\star \in \mathbb{R}^{n\times m}$ *which has rank* r*.*

Assumption 6.3.2. $f(\mathbf{X})$ *is* $(2r, 4r)$*-restricted strongly convex and smooth, i.e., for any* $n \times m$ *matrices* $\mathbf{X}, \mathbf{D}$ *with* $\mathrm{rank}(\mathbf{X}) \leq 2r$ *and* $\mathrm{rank}(\mathbf{D}) \leq 4r$*, the Hessian of* $f(\mathbf{X})$ *satisfies*

$$a\,\|\mathbf{D}\|_F^2 \leq [\nabla^2 f(\mathbf{X})](\mathbf{D}, \mathbf{D}) \leq b\,\|\mathbf{D}\|_F^2 \tag{6.6}$$

for some positive a *and* b*.*

[18] As mentioned previously, a similar notion of a revised robust strict saddle property has also recently been utilized in [126].

Assumption 6.3.1 is equivalent to the existence of a rank r $\mathbf{X}^\star$ such that $\nabla f(\mathbf{X}^\star) = \mathbf{0}$, which is very mild and holds in many matrix inverse problems including matrix sensing [65], matrix completion [177] and 1-bit matrix completion [132], where the unknown matrix to be recovered is a critical point of f.

Assumption 6.3.2 is also utilized in [109, Conditions 5.3 and 5.4] and [8], where weighted low-rank matrix factorization and a set of matrix inverse problems are proved to satisfy the $(2r, 4r)$-restricted strong convexity and smoothness condition (6.6). We discuss matrix sensing as a typical example satisfying this assumption in Section 6.3.3.

Combining Assumption 6.3.1 and Assumption 6.3.2, we have that $\mathbf{X}^\star$ is the unique global minimum of (6.1).

Proposition 6.3.1. *Suppose $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (6.6) *with positive a and b. Assume $\mathbf{X}^\star$ is a critical point of $f(\mathbf{X})$ with $\operatorname{rank}(\mathbf{X}^\star) = r$. Then $\mathbf{X}^\star$ is the global minimum of* (6.1)*, i.e.,*

$$f(\mathbf{X}^\star) \leq f(\mathbf{X}),\ \forall\, \mathbf{X} \in \mathbb{R}^{n\times m}, \operatorname{rank}(\mathbf{X}) \leq r$$

and the equality holds only at $\mathbf{X} = \mathbf{X}^\star$.

The proof of Proposition 6.3.1 is given in Appendix E.3. We note that Proposition 6.3.1 guarantees that $\mathbf{X}^\star$ is the unique global minimum of (6.1) and it is expected that solving the factorized problem (6.9) also gives $\mathbf{X}^\star$. Proposition 6.3.1 differs from [8] in that it only requires $\mathbf{X}^\star$ as a critical point, while [8] needs $\mathbf{X}^\star$ as a global minimum of f.

Before presenting the main result, we note that if f satisfies (6.6) with positive a and b and we rescale f as $f' = \frac{2}{a+b} f$, then f' satisfies

$$\frac{2a}{a+b}\|\mathbf{D}\|_F^2 \leq [\nabla^2 f'(\mathbf{X})](\mathbf{D},\mathbf{D}) \leq \frac{2b}{a+b}\|\mathbf{D}\|_F^2 .$$

It is clear that f and f' have the same optimization geometry (despite the scaling difference). Let $a' = \frac{2a}{a+b} = 1 - c$ and $b' = \frac{2a}{a+b} = 1 + c$ with $c = \frac{b-a}{a+b}$. We have $0 < a' \leq 1 \leq b'$ and $a' + b' = 2$. Thus, throughout the chapter and without the generality, we assume

$$a = 1 - c,\ b = 1 + c,\ c \in [0, 1). \tag{6.7}$$

Now let $\mathbf{X}^\star = \mathbf{\Phi\Sigma\Psi}^\top = \sum_{i=1}^r \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^\top$ be a reduced SVD of $\mathbf{X}^\star$, where $\mathbf{\Sigma}$ is a diagonal matrix with $\sigma_1 \geq \cdots \geq \sigma_r$ along its diagonal. Denote

$$\mathbf{U}^\star = \mathbf{\Phi\Sigma}^{1/2}\mathbf{R}, \mathbf{V}^\star = \mathbf{\Psi\Sigma}^{1/2}\mathbf{R} \tag{6.8}$$

for any $\mathbf{R} \in \mathcal{O}_r$. We first introduce the following ways to stack $\mathbf{U}$ and $\mathbf{V}$ together that are widely used through the chapter:

$$\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}, \quad \widehat{\mathbf{W}} = \begin{bmatrix} \mathbf{U} \\ -\mathbf{V} \end{bmatrix}, \mathbf{W}^\star = \begin{bmatrix} \mathbf{U}^\star \\ \mathbf{V}^\star \end{bmatrix}, \quad \widehat{\mathbf{W}}^\star = \begin{bmatrix} \mathbf{U}^\star \\ -\mathbf{V}^\star \end{bmatrix}.$$

Before moving on, we note that for any solution $(\mathbf{U}, \mathbf{V})$ to (6.2), $(\mathbf{U}\mathbf{R}_1, \mathbf{V}\mathbf{R}_2)$ is also a solution to (6.2) for any $\mathbf{R}_1, \mathbf{R}_2 \in \mathbb{R}^{r\times r}$ such that $\mathbf{U}\mathbf{R}_1\mathbf{R}_2^\top\mathbf{V}^\top = \mathbf{U}\mathbf{V}^\top$. As an extreme example, $\mathbf{R}_1 = c\mathbf{I}$ and $\mathbf{R}_2 = \frac{1}{c}\mathbf{I}$ where c can be arbitrarily large. In order to address this ambiguity (i.e., to reduce the search space of $\mathbf{W}$ for (6.3)), we utilize the trick in [8, 102, 106, 109] by introducing a regularizer ρ and turn to solve the following problem

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}, \mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \; G(\mathbf{W}) := h(\mathbf{W}) + \rho(\mathbf{W}), \tag{6.9}$$

where

$$\rho(\mathbf{W}) := \frac{\mu}{4}\left\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right\|_F^2.$$

We remark that $\mathbf{W}^\star$ is still a global minimizer of the factored problem (E.1) since both the first term and $\rho(\mathbf{W})$ achieve their global minimum at $\mathbf{W}^\star$. The regularizer $\rho(\mathbf{W})$ is applied to force the difference between the Gram matrices of $\mathbf{U}$ and $\mathbf{V}$ as small as possible. The global minimum of $\rho(\mathbf{W})$ is 0, which is achieved when $\mathbf{U}$ and $\mathbf{V}$ have the same Gram matrices, i.e., when $\mathbf{W}$ belongs to

$$\mathcal{E} := \left\{\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V} = \mathbf{0}\right\}. \tag{6.10}$$

Informally, we can view (6.9) as finding a point from $\mathcal{E}$ that also minimizes the first term in (6.9). This is rigorously established in the following result which reveals that any critical point $\mathbf{W}$ of $g(\mathbf{W})$ belongs to $\mathcal{E}$ (that is $\mathbf{U}$ and $\mathbf{V}$ are balanced factors of their product $\mathbf{U}\mathbf{V}^\top$) for any $\mu > 0$.

Lemma 6.3.1. *[8, Theorem 3] Suppose $G(\mathbf{W})$ is defined as in* (6.9) *with $\mu > 0$. Then any critical point $\mathbf{W}$ of $G(\mathbf{W})$ belongs to $\mathcal{E}$, i.e.,*

$$\nabla G(\mathbf{W}) = \mathbf{0} \quad \Rightarrow \quad \mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}. \tag{6.11}$$

For completeness, we include the proof of Lemma 6.3.1 in Appendix E.4.

6.3.2 Global Geometry For General Low-Rank Optimization

We now characterize the global optimization geometry of the factored problem (6.9). As explained in Section 6.2 that $G(\mathbf{W})$ is invariant under the matrices $\mathbf{R} \in \mathcal{O}_r$, we first recall the discussions in Section 6.2 about the revised robust strict saddle property for the invariant functions. To that end, we follow the notion of the distance between

equivalent classes for invariant functions defined in (6.4) and define the distance between $\mathbf{W}_1$ and $\mathbf{W}_2$ as follows

$$\begin{aligned} \operatorname{dist}(\mathbf{W}_1, \mathbf{W}_2) :&= \min_{\mathbf{R}_1 \in \mathcal{O}_r, \mathbf{R}_2 \in \mathcal{O}_r} \|\mathbf{W}_1 \mathbf{R}_1 - \mathbf{W}_2 \mathbf{R}_2\|_F \\ &= \min_{\mathbf{R} \in \mathcal{O}_r} \|\mathbf{W}_1 - \mathbf{W}_2 \mathbf{R}\|_F . \end{aligned} \tag{6.12}$$

For convenience, we also denote the best rotation matrix $\mathbf{R}$ so that $\|\mathbf{W}_1 - \mathbf{W}_2\mathbf{R}\|_F$ achieves its minimum by $\mathbf{R}(\mathbf{W}_1, \mathbf{W}_2)$, i.e.,

$$\mathbf{R}(\mathbf{W}_1, \mathbf{W}_2) := \arg \min_{\mathbf{R}' \in \mathcal{O}_r} \|\mathbf{W}_1 - \mathbf{W}_2 \mathbf{R}'\|_F , \tag{6.13}$$

which is also known as the orthogonal Procrustes problem [147]. The solution to the above minimization problem is characterized by the following lemma.

Lemma 6.3.2. *[147] Let $\mathbf{W}_2^\top \mathbf{W}_1 = \mathbf{L}\mathbf{S}\mathbf{P}^\top$ be an SVD of $\mathbf{W}_2^\top \mathbf{W}_1$. An optimal solution for the orthogonal Procrustes problem* (6.13) *is given by*

$$\mathbf{R}(\mathbf{W}_1, \mathbf{W}_2) = \mathbf{L}\mathbf{P}^\top.$$

Moreover, we have

$$\begin{aligned} \mathbf{W}_1^\top \mathbf{W}_2 \mathbf{R}(\mathbf{W}_1, \mathbf{W}_2) &= \left(\mathbf{W}_2 \mathbf{R}(\mathbf{W}_1, \mathbf{W}_2)\right)^\top \mathbf{W}_1 \\ &= \mathbf{P}\mathbf{S}\mathbf{P}^\top \succeq \mathbf{0}. \end{aligned}$$

To ease the notation, we drop $\mathbf{W}_1$ and $\mathbf{W}_2$ in $\mathbf{R}(\mathbf{W}_1, \mathbf{W}_2)$ and rewrite $\mathbf{R}$ instead of $\mathbf{R}(\mathbf{W}_1, \mathbf{W}_2)$ when they ($\mathbf{W}_1$ and $\mathbf{W}_2$) are clear from the context. Now we are well equipped to present the robust strict saddle property for $G(\mathbf{W})$ in the following result.

Theorem 6.3.1. *Define the following regions*

$$\mathcal{R}_1 := \left\{ \mathbf{W} : \operatorname{dist}(\mathbf{W}, \mathbf{W}^\star) \le \sigma_r^{1/2}(\mathbf{X}^\star) \right\},$$

$$\mathcal{R}_2 := \left\{ \mathbf{W} : \sigma_r(\mathbf{W}) \le \sqrt{\frac{1}{2}} \sigma_r^{1/2}(\mathbf{X}^\star), \|\mathbf{W}\mathbf{W}^\top\|_F \le \frac{20}{19} \|\mathbf{W}^\star \mathbf{W}^{\star\top}\|_F \right\},$$

$$\mathcal{R}_3' := \left\{ \mathbf{W} : \operatorname{dist}(\mathbf{W}, \mathbf{W}^\star) > \sigma_r^{1/2}(\mathbf{X}^\star), \|\mathbf{W}\| \le \frac{20}{19} \|\mathbf{W}^\star\|, \sigma_r(\mathbf{W}) > \sqrt{\frac{1}{2}} \sigma_r^{1/2}(\mathbf{X}^\star), \|\mathbf{W}\mathbf{W}^\top\|_F \le \frac{20}{19} \|\mathbf{W}^\star \mathbf{W}^{\star\top}\|_F \right\},$$

$$\mathcal{R}_3'' := \left\{ \mathbf{W} : \|\mathbf{W}\| > \frac{20}{19}\|\mathbf{W}^\star\| = \frac{20}{19}\sqrt{2}\|\mathbf{X}^\star\|^{1/2}, \|\mathbf{W}\mathbf{W}^\top\|_F \le \frac{10}{9}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F \right\},$$

$$\mathcal{R}_3''' := \left\{ \mathbf{W} : \|\mathbf{W}\mathbf{W}^\top\|_F > \frac{10}{9}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F = \frac{20}{9}\|\mathbf{X}^\star\|_F \right\}.$$

Let $G(\mathbf{W})$ be defined as in (6.9) *with $\mu = \frac{1}{2}$. Suppose $f(\mathbf{X})$ has a critical point $\mathbf{X}^\star \in \mathbb{R}^{n\times m}$ of rank r and satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (6.6) *with positive constants $a = 1 - c, b = 1 + c$ and*

$$c \lesssim \frac{\sigma_r^{3/2}(\mathbf{X}^\star)}{\|\mathbf{X}^\star\|_F \|\mathbf{X}^\star\|^{1/2}}. \tag{6.14}$$

Then $G(\mathbf{W})$ has the following robust strict saddle property:

1. *For any $\mathbf{W} \in \mathcal{R}_1$, $G(\mathbf{W})$ satisfies the local regularity condition:*

$$\langle \nabla G(\mathbf{W}), \mathbf{W} - \mathbf{W}^\star \rangle \gtrsim \sigma_r(\mathbf{X}^\star) \operatorname{dist}^2(\mathbf{W}, \mathbf{W}^\star) + \frac{1}{\|\mathbf{X}^\star\|} \|\nabla G(\mathbf{W})\|_F^2, \tag{6.15}$$

   where $\operatorname{dist}(\mathbf{W}, \mathbf{W}^\star)$ and $\mathbf{R}$ are defined in (6.12) *and* (6.13)*, respectively.*

2. *For any $\mathbf{W} \in \mathcal{R}_2$, $G(\mathbf{W})$ has a directional negative curvature, i.e.,*

$$\lambda_{\min}\left(\nabla^2 G(\mathbf{W})\right) \lesssim -\sigma_r(\mathbf{X}^\star). \tag{6.16}$$

3. *For any $\mathbf{W} \in \mathcal{R}_3 = \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$, $G(\mathbf{W})$ has large gradient descent:*

$$\|\nabla G(\mathbf{W})\|_F \gtrsim \sigma_r^{3/2}(\mathbf{X}^\star), \quad \forall\, \mathbf{W} \in \mathcal{R}_3'; \tag{6.17}$$
$$\|\nabla G(\mathbf{W})\|_F \gtrsim \|\mathbf{W}\|^3, \quad \forall\, \mathbf{W} \in \mathcal{R}_3''; \tag{6.18}$$
$$\|\nabla G(\mathbf{W})\|_F \gtrsim \sigma_r(\mathbf{X}^\star) \left(\|\mathbf{W}\mathbf{W}^\top\|_F\right)^{1/2}, \quad \forall\, \mathbf{W} \in \mathcal{R}_3'''. \tag{6.19}$$

The proof of this result is given in Appendix E.12. The main proof strategy is to utilize Assumption 6.3.1 and Assumption 6.3.2 about the function f to control the deviation between the gradient (and the Hessian) of the general low-rank optimization (6.9) and the counterpart of the matrix factorization problem so that the landscape of the general low-rank optimization (6.9) has a similar geometry property. To that end, in Appendix E.1, we provide a comprehensive geometric analysis for the matrix factorization problem (6.3). The reason for choosing $\mu = \frac{1}{2}$ is also discussed in Appendix E.1.6. We note that the results in Appendix E.1 are also of independent interest, as we show that the objective function in (6.3) obeys the strict saddle property and has no spurious local minima not only for exact-parameterization

($r = \text{rank}(\mathbf{X}^\star)$), but also for over-parameterization ($r > \text{rank}(\mathbf{X}^\star)$) and under-parameterization ($r < \text{rank}(\mathbf{X}^\star)$). Several remarks follow.

Remark 6.3.1. Note that

$$\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' \supseteq \left\{ \mathbf{W} : \|\mathbf{W}\| \leq \frac{20}{19}\|\mathbf{W}^\star\|_F, \|\mathbf{W}\mathbf{W}^\top\|_F \leq \frac{10}{9}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F \right\},$$

which further implies

$$\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' \cup \mathcal{R}_3'' \supseteq \{\mathbf{W} : \|\mathbf{W}\mathbf{W}^\top\|_F \leq \frac{10}{9}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F\}.$$

Thus, we conclude that $\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3''' = \mathbb{R}^{(n+m)\times r}$. Now the convergence analysis of the stochastic gradient descent algorithm in [114, 126] for the robust strict saddle functions also holds for $G(\mathbf{W})$.

Remark 6.3.2. The constants involved in Theorem 6.3.1 can be found in Appendix E.12 through the proof. Theorem 6.3.1 states that the objective function for the general low-rank optimization (6.9) also satisfies the robust strict saddle property when (6.14) holds. The requirement for c in (6.14) can be weakened to ensure the properties of $g(\mathbf{W})$ are preserved for $G(\mathbf{W})$ in some regions. For example, the local regularity condition (6.15) holds when

$$c \leq \frac{1}{50}$$

which is independent of $\mathbf{X}^\star$. With the analysis of the global geometric structure in $G(\mathbf{W})$, Theorem 6.3.1 ensures that many local search algorithms can converge to $\mathbf{X}^\star$ (which is the the global minimum of (6.1) as guaranteed by Proposition 6.3.1) with random initialization. In particular, stochastic gradient descent when applied to the matrix sensing problem (6.22) is guaranteed to find the global minimum $\mathbf{X}^\star$ in polynomial time.

Remark 6.3.3. Local (rather than global) geometry results for the general low-rank optimization (6.9) are also covered in [8], which only characterizes the geometry at all the critical points. Instead, Theorem 6.3.1 characterizes the global geometry for general low-rank optimization (6.9). Because the analysis is different, the proof strategy for Theorem 6.3.1 is also very different than that of [8]. Since [8] only considers local geometry, the result in [8] requires $c \leq 0.2$, which is slightly less restrictive than the one in (6.14).

Remark 6.3.4. To explain the necessity of the requirement on the constants a and b in (6.14), we utilize the symmetric weighted PCA problem (so that we can visualize the landscape of the factored problem in Figure 6.1) as an example where the objective function is

$$f(\mathbf{X}) = \frac{1}{2}\|\mathbf{\Omega} \odot (\mathbf{X} - \mathbf{X}^\star)\|_F^2, \tag{6.20}$$

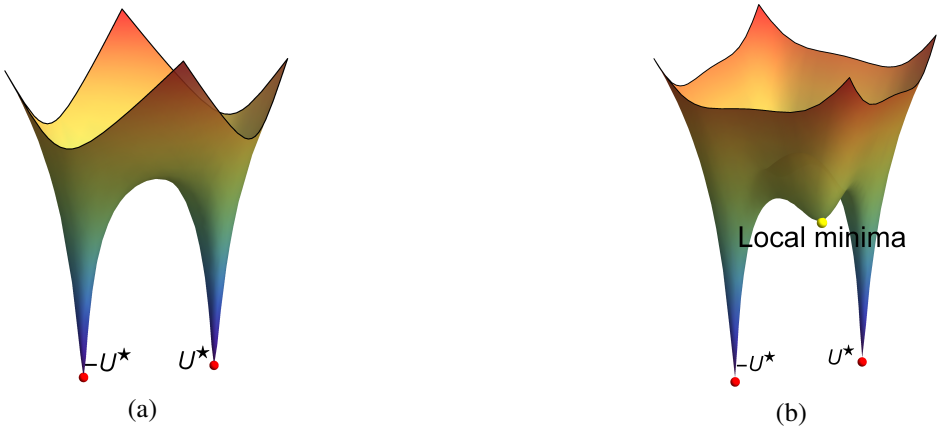
where $\mathbf{\Omega} \in \mathbb{R}^{n\times n}$ contains positive entries. The Hessian quadratic form for $f(\mathbf{X})$ is given by $[\nabla^2 f(\mathbf{X})](\mathbf{D}, \mathbf{D}) = \|\mathbf{\Omega} \odot \mathbf{D}\|_F^2$ for any $\mathbf{D} \in \mathbb{R}^{n\times n}$. Thus, we have

$$\min_{ij} |\mathbf{\Omega}[i,j]|^2 \leq \frac{[\nabla^2 f(\mathbf{X})](\mathbf{D},\mathbf{D})}{\|\mathbf{D}\|_F^2} \leq \max_{ij} |\mathbf{\Omega}[i,j]|^2.$$

Comparing with (6.6), we see that f satisfies the restricted strong convexity and smoothness conditions with the constants $a = \min_{ij} |\mathbf{\Omega}[i,j]|^2$ and $b = \max_{ij} |\mathbf{\Omega}[i,j]|^2$. In this case, we also note that if each entry W_{ij} is nonzero (i.e., $\min_{ij} |\mathbf{\Omega}[i,j]|^2 > 0$), the function $f(\mathbf{X})$ is strongly convex, rather than only restrictively strongly convex, implying that (6.20) has a unique optimal solution $\mathbf{X}^\star$. By applying the factorization approach, we get the factored objective function

$$h(\mathbf{U}) = \frac{1}{2}\|\mathbf{\Omega} \odot (\mathbf{U}\mathbf{U}^\top - \mathbf{X}^\star)\|_F^2. \tag{6.21}$$

To illustrate the necessity of the requirement on the constants a and b as in (6.14) so that the factored problem (6.21) has no spurious local minima and obeys the robust strict saddle property, we set $\mathbf{X}^\star = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ which is a rank-1 matrix and can be factorized as $\mathbf{X}^\star = \mathbf{U}^\star\mathbf{U}^{\star\top}$ with $\mathbf{U}^\star = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. We then plot the landscapes of the factored objective function $h(\mathbf{U})$ with $\mathbf{\Omega} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $\begin{bmatrix} 8 & 1 \\ 1 & 8 \end{bmatrix}$ in Figure 6.1. We observe from Figure 6.1 that as long as the elements in $\mathbf{\Omega}$ have a small dynamic range (which corresponds to a small b/a), $h(\mathbf{U})$ has no spurious local minima, but if the elements in $\mathbf{\Omega}$ have a large dynamic range (which corresponds to a large b/a), spurious local minima can appear in $h(\mathbf{U})$.



(a) (b)

Figure 6.1: Landscapes of $h(\mathbf{U})$ in (6.21) with $\mathbf{X}^\star = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and (a) $\mathbf{\Omega} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$; (b) $\mathbf{\Omega} = \begin{bmatrix} 8 & 1 \\ 1 & 8 \end{bmatrix}$.

Remark 6.3.5. The global geometry of low-rank matrix recovery but with analysis customized to linear measurements and quadratic loss functions is also covered in [101, 104]. Since Theorem 6.3.1 only requires the $(2r, 4r)$-restricted strong convexity and smoothness property (6.6), aside from low-rank matrix recovery [189], it can also be applied to many other low-rank matrix optimization problems [161] which do not necessarily involve quadratic loss functions. Typical examples include 1-bit matrix completion [132, 164] and Poisson principal component analysis (PCA) [165]. We refer to [8] for more discussion on this issue. In next section, we consider a stylized application of Theorem 6.3.1 in matrix sensing and compare it with the result in [104].

6.3.3 Stylized Application: Matrix Sensing

In this section, we extend the previous geometric analysis to the matrix sensing problem

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}}\ G(\mathbf{W}) := \frac{1}{2}\left\|\mathcal{A}\left(\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\right)\right\|_2^2 + \rho(\mathbf{W}), \tag{6.22}$$

where $\mathcal{A}: \mathbb{R}^{n\times m} \to \mathbb{R}^p$ is a known linear measurement operator and $\mathbf{X}^\star$ is the unknown rank r matrix to be recovered. In this case, we have

$$f(\mathbf{X}) = \frac{1}{2}\left\|\mathcal{A}\left(\mathbf{X} - \mathbf{X}^\star\right)\right\|_2^2.$$

The derivative of $f(\mathbf{X})$ at $\mathbf{X}^\star$ is

$$\nabla f(\mathbf{X}^\star) = \mathcal{A}^*\mathcal{A}(\mathbf{X}^\star - \mathbf{X}^\star) = \mathbf{0},$$

which implies that $f(\mathbf{X})$ satisfies Assumption 6.3.1. The Hessian quadrature form $\nabla^2 f(\mathbf{X})[\mathbf{D}, \mathbf{D}]$ for any $n \times m$ matrices $\mathbf{X}$ and $\mathbf{D}$ is given by

$$\nabla^2 f(\mathbf{X})[\mathbf{D}, \mathbf{D}] = \left\|\mathcal{A}(\mathbf{D})\right\|^2.$$

The following matrix Restricted Isometry Property (RIP) serves as a way to link the low-rank matrix factorization problem (E.1) with the matrix sensing problem (6.22) and certifies $f(\mathbf{X})$ satisfying Assumption 6.3.2.

Definition 6.3.1 (Restricted Isometry Property (RIP) [65, 190])**.** *The map $\mathcal{A}: \mathbb{R}^{n\times m} \to \mathbb{R}^p$ satisfies the r-RIP with constant δ_r if*[19]

$$(1-\delta_r)\left\|\mathbf{D}\right\|_F^2 \le \left\|\mathcal{A}(\mathbf{D})\right\|^2 \le (1+\delta_r)\left\|\mathbf{D}\right\|_F^2 \tag{6.23}$$

holds for any $n \times m$ matrix $\mathbf{D}$ with $\operatorname{rank}(\mathbf{D}) \le r$.

[19] By abuse of notation, we adopt the conventional notation δ_r for the RIP constant. The subscript r can be used to distinguish the RIP constant δ_r from δ which is used as a small constant in Section 6.2.

If $\mathcal{A}$ satisfies the $4r$-restricted isometry property with constant δ_{4r}, then $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (6.6) with constants $a = 1 - \delta_{4r}$ and $b = 1 + \delta_{4r}$ since

$$\begin{aligned}(1 - \delta_{4r}) \|\mathbf{D}\|_F^2 \leq \nabla^2 f(\mathbf{X})[\mathbf{D}, \mathbf{D}] = \|\mathcal{A}(\mathbf{D})\|^2 \\ \leq (1 + \delta_{4r}) \|\mathbf{D}\|_F^2\end{aligned} \tag{6.24}$$

for any rank-$4r$ matrix $\mathbf{D}$. Comparing (6.24) with (6.6), we note that the RIP is stronger than the restricted strong convexity and smoothness property (6.6) as the RIP gives that (6.24) holds for all $n \times m$ matrices $\mathbf{X}$, while Assumption 6.3.2 only requires that (6.6) holds for all rank-$2r$ matrices.

Now, applying Theorem 6.3.1, we obtain a similar geometric guarantee to Theorem 6.3.1 for the matrix sensing problem (6.22) when $\mathcal{A}$ satisfies the RIP.

Corollary 6.3.1. *Let $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3', \mathcal{R}_3'', \mathcal{R}_3'''$ be the regions as defined in Theorem E.11.1. Let $G(\mathbf{W})$ be defined as in* (6.22) *with $\mu = \frac{1}{2}$ and $\mathcal{A}$ satisfying the $4r$-RIP with*

$$\delta_{4r} \lesssim \frac{\sigma_r^{3/2}(\mathbf{X}^\star)}{\|\mathbf{X}^\star\|_F \|\mathbf{X}^\star\|^{1/2}}. \tag{6.25}$$

Then $G(\mathbf{W})$ has the following robust strict saddle property:

1. *For any $\mathbf{W} \in \mathcal{R}_1$, $G(\mathbf{W})$ satisfies the local regularity condition:*

$$\langle \nabla G(\mathbf{W}), \mathbf{W} - \mathbf{W}^\star \rangle \gtrsim \sigma_r(\mathbf{X}^\star) \operatorname{dist}^2(\mathbf{W}, \mathbf{W}^\star) + \frac{1}{\|\mathbf{X}^\star\|} \|\nabla G(\mathbf{W})\|_F^2, \tag{6.26}$$

   where $\operatorname{dist}(\mathbf{W}, \mathbf{W}^\star)$ and $\mathbf{R}$ are defined in (6.12) *and* (6.13)*, respectively.*

2. *For any $\mathbf{W} \in \mathcal{R}_2$, $G(\mathbf{W})$ has a directional negative curvature, i.e.,*

$$\lambda_{\min}\left(\nabla^2 G(\mathbf{W})\right) \lesssim -\sigma_r(\mathbf{X}^\star).$$

3. *For any $\mathbf{W} \in \mathcal{R}_3 = \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$, $G(\mathbf{W})$ has large gradient descent:*

$$\begin{aligned}&\|\nabla G(\mathbf{W})\|_F \gtrsim \sigma_r^{3/2}(\mathbf{X}^\star), \quad \forall\, \mathbf{W} \in \mathcal{R}_3'; \\ &\|\nabla G(\mathbf{W})\|_F \gtrsim \|\mathbf{W}\|^3, \quad \forall\, \mathbf{W} \in \mathcal{R}_3''; \\ &\|\nabla G(\mathbf{W})\|_F \gtrsim \left\|\mathbf{W}\mathbf{W}^\top\right\|_F^{3/2}, \quad \forall\, \mathbf{W} \in \mathcal{R}_3'''.\end{aligned}$$

Remark 6.3.6. The constants involved in Corollary 6.3.1 are the same as those in Theorem 6.3.1 and can be found in Appendix E.12 through the proof. Similarly, the requirement for δ_{4r} in (6.25) can be weakened to ensure the properties of $g(\mathbf{W})$ are preserved for $G(\mathbf{W})$ in some regions. For example, the local regularity condition (6.26) holds when

$$\delta_{4r} \leq \frac{1}{50}$$

which is independent of $\mathbf{X}^\star$. Note that Tu et al. [102, Section 5.4, (5.15)] provided a similar regularity condition. However, the result there requires $\delta_{6r} \leq \frac{1}{25}$ and $\operatorname{dist}(\mathbf{W}, \mathbf{W}^\star) \leq \frac{1}{2\sqrt{2}}\sigma_r(\mathbf{X}^\star)$ which defines a smaller region than $\mathcal{R}_1$. Based on this local regularity condition, Tu et al. [102] showed that gradient descent with a good initialization (which is close enough to $\mathbf{W}^\star$) converges to the unknown matrix $\mathbf{W}^\star$ (and hence $\mathbf{X}^\star$). With the analysis of the global geometric structure in $G(\mathbf{W})$, Corollary 6.3.1 ensures that many local search algorithms can find the unknown matrix $\mathbf{X}^\star$ in polynomial time.

Remark 6.3.7. A Gaussian $\mathcal{A}$ will have the RIP with high probability when the number of measurements p is comparable to the number of degrees of freedom in an $n \times m$ matrix with rank r. By Gaussian $\mathcal{A}$ we mean the ℓ-th element in $\mathbf{y} = \mathcal{A}(\mathbf{X})$, y_ℓ, is given by

$$y_\ell = \langle \mathbf{X}, \mathbf{A}_\ell \rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{X}[i,j]\mathbf{A}_\ell[i,j],$$

where the entries of each $n \times m$ matrix $\mathbf{A}_\ell$ are independent and identically distributed normal random variables with zero mean and variance $\frac{1}{p}$. Specifically, a Gaussian $\mathcal{A}$ satisfies (6.23) with high probability when [65, 124, 189]

$$p \gtrsim r(n+m)\frac{1}{\delta_r^2}.$$

Now utilizing the inequality $\|\mathbf{X}^\star\|_F \leq \sqrt{r}\|\mathbf{X}^\star\|$ for (6.14), we conclude that in the case of Gaussian measurements, the robust strict saddle property is preserved for the matrix sensing problem with high probability when the number of measurements exceeds a constant times $(n+m)r^2\kappa(\mathbf{X}^\star)^3$ where $\kappa(\mathbf{X}^\star) = \frac{\sigma_1(\mathbf{X}^\star)}{\sigma_r(\mathbf{X}^\star)}$. This further implies that, when applying the stochastic gradient descent algorithm to the matrix sensing problem (6.22) with Gaussian measurements, we are guaranteed to find the unknown matrix $\mathbf{X}^\star$ in polynomial time with high probability when

$$p \gtrsim (n+m)r^2\kappa(\mathbf{X}^\star)^3. \tag{6.27}$$

When $\mathbf{X}^\star$ is an $n \times n$ PSD matrix, Li et al. [104] showed that the corresponding matrix sensing problem with Gaussian measurements has similar global geometry to the low-rank PSD matrix factorization problem when the number of measurements

$$p \gtrsim nr^2\frac{\sigma_1^4(\mathbf{X}^\star)}{\sigma_r^2(\mathbf{X}^\star)}. \tag{6.28}$$

Comparing (6.27) with (6.28), we find both results for the number of measurements needed depend similarly on the rank r, but slightly differently on the spectrum of $\mathbf{X}^\star$. We finally remark that the sampling complexity in (6.27) is $O((n+m)r^2)$, which is slightly larger than the information theoretically optimal bound $O((n+m)r)$ for matrix

sensing. This is because Corollary 6.3.1 is a direct consequence of Theorem 6.3.1 in which we directly characterize the landscapes of the objective functions in the whole space by combining the results for matrix factorization in Appendix E.1 and the restricted strong convexity and smoothness condition. We believe this mismatch is an artifact of our proof strategy and could be mitigated by a different approach, like utilizing the properties of quadratic loss functions [101]. If one desires only to characterize the geometry for critical points, then $O((n+m)r)$ measurements are enough to ensure the strict saddle property and lack of spurious local minima for matrix sensing [8, 106].

CHAPTER 7

THE GEOMETRY OF EQUALITY-CONSTRAINED GLOBAL CONSENSUS PROBLEMS

A variety of unconstrained nonconvex optimization problems have been shown to have benign geometric landscapes that satisfy the strict saddle property and have no spurious local minima. In this work[20], we present a general result relating the geometry of an unconstrained centralized problem to its equality-constrained distributed extension. It follows that many global consensus problems inherit the benign geometry of their original centralized counterpart. Taking advantage of this fact, we demonstrate the favorable performance of the Gradient ADMM algorithm on a distributed low-rank matrix approximation problem.

7.1 Introduction

With an abundance of data, the scale of machine learning problems continues to grow. Consequently, nonconvex optimization problems have received growing attention as alternatives to convex approaches for solving machine learning problems [124, 191–193]. Algorithms for solving nonconvex problems can offer reduced memory usage and computational complexity compared to their convex counterparts, see, e.g. [119, 194]. However, the potential for undesirable features in the nonconvex landscape (spurious local minima [135–137], degenerate saddle points [137, 195], etc.) raises questions about these algorithms' convergence to optimal points.

Recent research has shown, though, that many machine learning problems—including a variety of low-rank matrix optimization problems—actually have a benign nonconvex landscape in which there are no spurious local minima and all saddle points are strict (non-degenerate) saddles at which the Hessian has at least one negative eigenvalue [6, 8, 93, 100, 101, 104, 125, 191–193, 196]. For such problems a variety of iterative algorithms—such as gradient descent with a random initialization—can exploit negative curvature directions to escape from strict saddle points and thus provably converge to a global minimizer [127].

To date, however, most of the results establishing benign geometric landscapes have been limited to *unconstrained* nonconvex problems [6, 8, 92, 93, 100, 101, 104, 125, 196]. Meanwhile, constraints can be important to consider, particularly when the size of a machine learning problem demands that computations or storage be *distributed* across some network [197, 198]. One way to ensure consensus among optimization variables in a distributed problem is via equality constraints across the network nodes. As one transitions from a centralized problem to a distributed one, a question arises of whether the distributed problem inherits the benign geometry of the centralized problem. Since there is a general lack of geometric analysis for constrained nonconvex problems, this question is essentially open.

[20] This is a joint work with Zhihui Zhu, Gongguo Tang and Michael B. Wakin [9].

In Section 7.2, we present a general result relating the geometry of a centralized problem to its distributed extension. This result establishes one-to-one correspondences of the first-order critical points, second-order critical points, and strict saddle points between the two problems. This is in spite of the fact that critical points have a distinctly different definition (in terms of the Lagrangian) for constrained optimization problems. In Section 7.3, we highlight one application of this theorem, in establishing an equivalence between geometric landscapes for broad classes of centralized problems and their distributed formulations as global consensus problems. We show that under certain conditions, every second-order critical point of the distributed problem corresponds to a global minimizer of the centralized problem. In Section 7.4, we discuss algorithmic aspects for solving equality-constrained distributed optimization problems. The recent GADMM algorithm [199] can be guaranteed under certain conditions to converge to a second-order critical point of an equality-constrained distributed optimization problem. Our theory establishes conditions under which this point will correspond to a global minimizer of the original centralized optimization problem. This guarantee is stronger than what appear in the literature for distributed gradient descent (DGD), a popular alternative algorithm for solving consensus problems. Existing DGD results show convergence either to stationary points (which are global if the problem is convex) [200–202], or to an arbitrarily small neighborhood of a second-order critical point with an appropriately small stepsize [203]. As a case study, in Section 7.5, we illustrate the performance of GADMM on a distributed low-rank matrix approximation in factored form.

7.2 Relating Unconstrained Geometry to Constrained Geometry

We present a general theorem that establishes an equivalence between the landscape of two types of optimization problems: one that is unconstrained, and one that involves additional variables but is constrained to an affine subspace, along which it has a certain equivalence to the first problem.

Theorem 7.2.1. *Consider two problems:*

- *Problem UC (unconstrained centralized):*

$$\min_{\mathbf{x}} c(\mathbf{x})$$

- *Problem ECD (equality-constrained distributed):*

$$\min_{\mathbf{x},\mathbf{y}} d(\mathbf{x},\mathbf{y}) \text{ subject to } \mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}=\mathbf{b}$$

where $d(\mathbf{x},\mathbf{y})$ satisfies $d(\mathbf{x},\mathbf{y})=c(\mathbf{x})$ when $\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}=\mathbf{b}$, and $\mathbf{B}$ is a square and invertible matrix.

Then $\mathbf{x}$ is a [first-order/second-order/strict saddle] critical point of Problem UC iff $(\mathbf{x},\mathbf{B}^{-1}(\mathbf{b}-\mathbf{A}\mathbf{x}))$ is a [first-order/second-order/strict saddle] point of Problem ECD.

This theorem has applications outside of distributed optimization, but we adopt the terminology "centralized" and "distributed" in the theorem above because the latter problem involves additional optimization variables beyond those

in the first, and we focus on applications in distributed optimization in this work. Now we prove Theorem 7.2.1 in the follows where the precise notions of [first-order/second-order/strict saddle] point are defined for both Problem UC and Problem ECD. Critical points of Problem ECD are defined in terms of the Lagrangian function for $d(\mathbf{x}, \mathbf{y})$.

Proof. The first-order critical points $\mathbf{x}$ of Problem UC are those that satisfy

$$\nabla_{\mathbf{x}} c(\mathbf{x}) = 0. \tag{7.1}$$

The second-order critical points of Problem UC additionally satisfy

$$\nabla^2_{\mathbf{x}} c(\mathbf{x}) \succeq 0, \tag{7.2}$$

and a first-order critical point is a strict saddle if it does not satisfy (7.2).

The critical points $(\mathbf{x}, \mathbf{y})$ of Problem ECD are defined through the Lagrangian function $\mathcal{L}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = d(\mathbf{x}, \mathbf{y}) - \boldsymbol{\lambda}^\top (\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} - \mathbf{b})$. The first-order critical points $(\mathbf{x}, \mathbf{y})$ of Problem ECD are those that satisfy the first-order optimality condition: $\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{b}$ and there exists $\boldsymbol{\lambda}$ such that

$$\nabla_{\left[\mathbf{x}; \mathbf{y}\right]} \mathcal{L}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = 0. \tag{7.3}$$

The second-order critical points of Problem ECD additionally satisfy the second-order optimality condition:

$$[\nabla^2_{\left[\mathbf{x}; \mathbf{y}\right]} \mathcal{L}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})](\mathbf{v}, \mathbf{v}) \geq 0 \; \forall \mathbf{v} \in \mathcal{T}, \tag{7.4}$$

where

$$\mathcal{T} = \{\mathbf{v} = \left[\mathbf{v}_x; \mathbf{v}_y\right] : \mathbf{A}\mathbf{v}_x + \mathbf{B}\mathbf{v}_y = 0\} = \begin{bmatrix} \mathbb{R}^n \\ -\mathbf{B}^{-1}\mathbf{A}(\mathbb{R}^n) \end{bmatrix} \tag{7.5}$$

is the tangent plane of the constraint set $\mathcal{F} = \{\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{b}\}$, where we have used the nonsingularity of $\mathbf{B}$. A first-order critical point is a strict saddle if it does not satisfy (7.4).

For convenience, define

$$h(\mathbf{x}, \mathbf{y}) := d(\mathbf{x}, \mathbf{y}) - c(\mathbf{x}), \tag{7.6}$$

and note that $h(\mathbf{x}, \mathbf{y}) = 0$ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{F}$. Note that $h(\mathbf{x}, \mathbf{y})$ has zero directional derivative and zero Hessian curvature along the tangent plane of $\mathcal{F}$. That is

$$\nabla_{\left[\mathbf{x}; \mathbf{y}\right]} h(\mathbf{x}, \mathbf{y})^\top \mathbf{v} = 0 \text{ and } [\nabla^2_{\left[\mathbf{x}; \mathbf{y}\right]} h(\mathbf{x}, \mathbf{y})](\mathbf{v}, \mathbf{v}) = 0 \tag{7.7}$$

for any $(\mathbf{x}, \mathbf{y}) \in \mathcal{F}$ and $\mathbf{v} \in \mathcal{T}$.

For any $\mathbf{x}$, let $\mathbf{y} = \mathbf{B}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x})$ and note that $(\mathbf{x}, \mathbf{y}) \in \mathcal{F}$. Moreover, (7.3) holds iff $\begin{bmatrix} \nabla_{\mathbf{x}} c(\mathbf{x}); 0 \end{bmatrix} \perp \mathcal{T}$ (due to (7.7)), which holds iff (7.1) holds (due to (7.5)). Similarly, we have that (7.4) holds iff $[\nabla^2_{\begin{bmatrix} \mathbf{x}; \mathbf{y} \end{bmatrix}} c(\mathbf{x})](\mathbf{v}, \mathbf{v}) \geq 0 \; \forall \mathbf{v} \in \mathcal{T}$ (due to (7.6) and (7.7)), which holds iff (7.2) holds (due to (7.5)). This completes the proof of the three types of equivalence between a critical point $\mathbf{x}$ of Problem UC and a critical point $(\mathbf{x}, \mathbf{B}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}))$ of Problem ECD. $\square$

7.3 Geometry of Global Consensus

Consider any unconstrained centralized optimization problem of the form

$$\underset{\mathbf{w},\{\mathbf{z}_j\}}{\text{minimize}} \left(\sum_{j=1}^{J} f_j(\mathbf{w}, \mathbf{z}_j) \right) + g(\mathbf{w}), \tag{7.8}$$

where first term in the objective function decouples into a sum of objectives f_j. One can distribute this problem across a network of $J+1$ nodes in a "star topology",[21] where J agents are connected to a central node. The resulting problem is known as a *global consensus problem* (see [199, (3)]) and can be posed as follows[22]:

$$\underset{\mathbf{w},\{\mathbf{z}_j\},\{\mathbf{w}^j\}}{\text{minimize}} \left(\sum_{j=1}^{J} f_j(\mathbf{w}^j, \mathbf{z}_j) \right) + g(\mathbf{w}) \text{ s.t. } \mathbf{w}^j = \mathbf{w} \; \forall j. \tag{7.9}$$

Here, $\mathbf{w}$ is the optimization variable at the central node, and $\mathbf{w}^j$ and $\mathbf{z}_j$ are the optimization variables at node j.

Unfortunately, relatively little is currently known about the geometric landscape of equality-constrained machine learning problems in the form of (7.9): Do they have spurious local minima? Do they satisfy the strict saddle property, or could they have degenerate saddle points?

However, insight into the geometry of problem (7.9) can be gained by applying Theorem 7.2.1. Problem (7.8) can be expressed in the form of Problem UC by taking[23] $\mathbf{x} = [\mathbf{w}; \mathbf{z}]$ with $\mathbf{z} = [\mathbf{z}_1; \cdots; \mathbf{z}_J]$ and $c(\mathbf{x}) = \sum_{j=1}^{J} f_j(\mathbf{w}, \mathbf{z}_j) + g(\mathbf{w})$, while problem (7.9) can be expressed in the form of Problem ECD by taking $\mathbf{x} = [\mathbf{w}; \mathbf{z}]$, $\mathbf{y} = [\mathbf{w}^1; \cdots; \mathbf{w}^J]$, $d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{J} f_j(\mathbf{w}^j, \mathbf{z}_j) + g(\mathbf{w})$,

$$\mathbf{A} = \begin{bmatrix} -\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ -\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}, \; \mathbf{B} = \begin{bmatrix} \mathbf{I} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}, \; \mathbf{b} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}. \tag{7.10}$$

We note that $\mathbf{B}$ (the identity matrix) is square and invertible. Under the constraint that $\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{b}$, which requires all $\mathbf{w}^j = \mathbf{w}$, we see that $d(\mathbf{x}, \mathbf{y}) = c(\mathbf{x})$. By applying Theorem 7.2.1, we obtain the following result.

Corollary 7.3.1. *$[\mathbf{w}; \mathbf{z}]$ is a [first-order/second-order/strict saddle] critical point of problem* (7.8) *iff $([\mathbf{w}; \mathbf{z}], [\mathbf{w}; \cdots; \mathbf{w}])$ is a [first-order/second-order/strict saddle] point of problem* (7.9)*. Moreover, if problem* (7.8) *satisfies the strict saddle property and has no spurious local minima, then for every second-order critical point $([\mathbf{w}; \mathbf{z}], [\mathbf{w}; \cdots; \mathbf{w}])$ of problem* (7.9)*, $[\mathbf{w}; \mathbf{z}]$ is a global minimizer of problem* (7.8)*.*

Corollary 7.3.1 allows one to borrow centralized geometric analysis for problem (7.8) to understand the landscape of the equality-constrained distributed problem (7.9).

[21] We remark that our results can also be applied to other network topologies, such as the series topology where $\mathbf{w}^j = \mathbf{w}^{j+1}$, $\forall j$ and $\mathbf{w}^J = \mathbf{w}$.
[22] Strictly speaking, our problem (7.9) is more general than [199, (3)] as (7.9) involves local variables $\{\mathbf{z}_j\}$ which are not constrained to be equal.
[23] To simplify notation, we use $[\mathbf{p}; \mathbf{q}]$ to represent $[\mathbf{p}^\top \; \mathbf{q}^\top]^\top$.

7.4 Gradient ADMM (GADMM) Algorithm

We briefly discuss algorithmic aspects for solving equality-constrained distributed optimization problems. The recent Gradient ADMM (GADMM) algorithm [199] can be guaranteed under certain conditions to converge to a second-order critical point of an equality-constrained distributed optimization problem. Corollary 7.3.1 establishes conditions under which this point will correspond to a global minimizer of the original centralized optimization problem.

As outlined in [199, (38)], GADMM is intended for problems that can be expressed as[24]

$$\underset{\mathbf{x},\mathbf{y}}{\text{minimize}}\, f(\mathbf{x}) + g(\mathbf{y}) \quad \text{subject to} \quad \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{b}. \tag{7.11}$$

The global consensus problem (7.9) is of this form; to see this, let $\mathbf{x} = [\mathbf{w}^1; \cdots; \mathbf{w}^J; \mathbf{z}]$, $\mathbf{y} = \mathbf{w}$, $f(\mathbf{x}) = \sum_{j=1}^{J} f_j(\mathbf{w}^j, \mathbf{z}_j)$, $g(\mathbf{y}) = g(\mathbf{w})$,

$$\mathbf{A} = \begin{bmatrix} -\mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & -\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} \mathbf{I} \\ \vdots \\ \mathbf{I} \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}. \tag{7.12}$$

In [199, Section 3.1], it is shown how GADMM can be applied to the global consensus problem (7.9), with the resulting iterations

$$\begin{aligned} \mathbf{w}^j(k+1) &= \mathbf{w}^j(k) - \frac{1}{\beta}\big(\nabla f_j(\mathbf{w}^j(k), \mathbf{z}_j(k)) \\ &\qquad + \boldsymbol{\lambda}_j(k) + \rho(\mathbf{w}^j(k) - \mathbf{w}(k))\big), \\ \mathbf{z}_j(k+1) &= \mathbf{z}_j(k) - \frac{1}{\beta}\nabla f_j(\mathbf{w}^j(k), \mathbf{z}_j(k)), \\ \mathbf{w}(k+1) &= \mathbf{w}(k) - \frac{1}{\beta}\big(\nabla g(\mathbf{w}_k) \\ &\qquad - \sum_{j=1}^{J}(\boldsymbol{\lambda}_j(k) + \rho(\mathbf{w}^j(k+1) - \mathbf{w}(k))\big), \\ \boldsymbol{\lambda}_j(k+1) &= \boldsymbol{\lambda}_j(k) + \rho(\mathbf{w}^j(k) - \mathbf{w}(k)). \end{aligned} \tag{7.13}$$

These iterations require communication only between the central node and each of the nodes $1, 2, \ldots, J$. We note that this is the reason that we utilize (7.12) instead of (7.10) when applying GADMM for solving (7.9) since the resulting algorithm (7.13) is more suitable for distributed implementation. On the other hand, the form (7.10) is mainly utilized for analyzing the landscape of (7.9) by invoking Theorem 7.2.1.

For the global consensus problem, under assumptions B1–B5 in [199], with the proper selection of parameters β and ρ, and with random initialization of $\mathbf{w}(0)$, $\{\mathbf{w}^j(0)\}$, $\{\mathbf{z}_j(0)\}$, and $\{\boldsymbol{\lambda}_j(0)\}$ it is shown [199, Theorem 3.1] that with probability one, GADMM will converge to a second-order critical point of (7.9). According to Corollary 7.3.1,

[24] The notations f and g are interchanged with respect to what appears in [199, (38)].

when problem (7.8) satisfies the strict saddle property and has no spurious local minima, this second-order critical point of (7.9) corresponds to a global minimizer of problem (7.8).

7.5 Application to Distributed Low-Rank Matrix Approximation

We now discuss our results in the context of distributed low-rank matrix approximation. Consider first the prototypical problem of finding, for a given a data matrix $\mathbf{Y} \in \mathbb{R}^{n \times m}$, a low-rank approximation by solving

$$\underset{\mathbf{X} \in \mathbb{R}^{n \times m}}{\text{minimize}} \, \|\mathbf{X} - \mathbf{Y}\|_F^2 + \mu \|\mathbf{X}\|_*. \tag{7.14}$$

Here, the nuclear norm penalty promotes low-rank structure in the approximation $\mathbf{X}$. Problem (7.14) is an unconstrained convex optimization problem in the matrix variable $\mathbf{X}$. It is natural to consider solving problem (7.14) in factored form, where we replace the optimization variable $\mathbf{X}$ with $\mathbf{U}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n \times r}$ and $\mathbf{V} \in \mathbb{R}^{m \times r}$ are tall matrices, and r is a parameter that must be set in advance (typically on the order of the rank r' expected of the optimal solution to (7.14)). Under this reparameterization, (7.14) becomes

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{V} \in \mathbb{R}^{m \times r}}{\text{minimize}} \, \|\mathbf{U}\mathbf{V}^\top - \mathbf{Y}\|_F^2 + \mu \|\mathbf{U}\mathbf{V}^\top\|_*. \tag{7.15}$$

One can solve this problem using local search algorithms such as gradient descent. Such algorithms do not require expensive SVDs, nor do they require explicit storage of the matrix $\mathbf{X}$.

Unfortunately, problem (7.15) is nonconvex in the optimization variables $(\mathbf{U}, \mathbf{V})$. We have studied [6] the geometric landscape of problem (7.15) with a minor modification to the objective function:

$$\underset{\mathbf{U}, \mathbf{V}}{\text{minimize}} \, \|\mathbf{U}\mathbf{V}^\top - \mathbf{Y}\|_F^2 + \frac{\mu}{2} \left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 \right). \tag{7.16}$$

Despite the change of the objective function, the global minimizers remain unchanged. That is, any $(\mathbf{U}, \mathbf{V})$ that minimize (7.16) are also a global minimizer of (7.15).

We have shown [6] that every critical point of problem (7.16) is either a global minimum or a strict saddle point. This implies that local search algorithms such as gradient descent can be applied to problem (7.16) and will converge to a global minimum of (7.16). As previously noted, this then coincides with a global minimum of the original objective function, (7.15). This favorable geometry of problem (7.16) holds under the condition that there exists a global minimizer of (7.14) having rank r' and that $r \geq r'$.

One can generalize the unconstrained centralized problem (7.16) to an equality-constrained distributed problem similar to the global consensus problems outlined in Section 7.3. Suppose the columns of the data matrix $\mathbf{Y}$ are distributed among J nodes/sensors. Without loss of generality, partition the columns of $\mathbf{Y}$ as $\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 & \cdots & \mathbf{Y}_J \end{bmatrix}$ where for $j \in \{1, 2, \ldots, J\}$, matrix $\mathbf{Y}_j$ (which is stored at node j) has size $n \times m_j$, and where $m = \sum_{j=1}^J m_j$. Par-

titioning $\mathbf{V}$ similarly as $\mathbf{V} = \begin{bmatrix} \mathbf{V}_1^\top & \mathbf{V}_2^\top & \cdots & \mathbf{V}_J^\top \end{bmatrix}^\top$, where $\mathbf{V}_j$ has size $m_j \times r$, we can write $\|\mathbf{U}\mathbf{V}^\top - \mathbf{Y}\|_F^2 = \sum_{j=1}^J \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2$. We use this fact to plug in for the term $\|\mathbf{U}\mathbf{V}^\top - \mathbf{Y}\|_F^2$ which appears in (7.16).

Suppose we introduce in problem (7.16) the optimization variables $\mathbf{U}^1, \ldots, \mathbf{U}^J \in \mathbb{R}^{n \times r}$ (all the same size as $\mathbf{U}$) and add an equality constraint to enforce consensus among these variables. We obtain the equality-constrained optimization problem

$$\begin{aligned} \underset{\mathbf{U}, \{\mathbf{V}_j\}, \{\mathbf{U}^j\}}{\text{minimize}} \left(\sum_{j=1}^J \|\mathbf{U}^j \mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2 + \frac{\mu}{2} \|\mathbf{V}_j\|_F^2 \right) \\ + \frac{\mu}{2} \|\mathbf{U}\|_F^2 \text{ subject to } \mathbf{U}^j = \mathbf{U}, \ \forall\, j, \end{aligned} \tag{7.17}$$

which has the form of global consensus problem appearing in (7.9) by taking $\mathbf{w} = \text{vec}(\mathbf{U}), \mathbf{z}_j = \text{vec}(\mathbf{V}_j)$, $\mathbf{w}^j = \text{vec}(\mathbf{U}^j)$, and defining $f_j(\mathbf{w}^j, \mathbf{z}_j)$, $g(\mathbf{w})$ in the natural resulting way. By applying Corollary 7.3.1, we obtain the following result.

Corollary 7.5.1. *$\mathbf{x}_{UC} = [vec(\mathbf{U}); vec(\mathbf{V}_1); \cdots; vec(\mathbf{V}_J)]$ is a [first-order/second-order/strict saddle] critical point of problem* (7.16) *iff $\mathbf{x}_{ECD} = (\mathbf{x}_{UC}, [vec(\mathbf{U}); \cdots; vec(\mathbf{U})])$ is a [first-order/second-order/strict saddle] critical point of problem* (7.17)*. Moreover, under the condition that there exists a global minimizer of* (7.14) *having rank r' and that $r \geq r'$, for every second-order critical point $\mathbf{x}_{ECD}$ of problem* (7.17)*, $\mathbf{x}_{UC}$ is a global minimizer of problem* (7.16)*.*

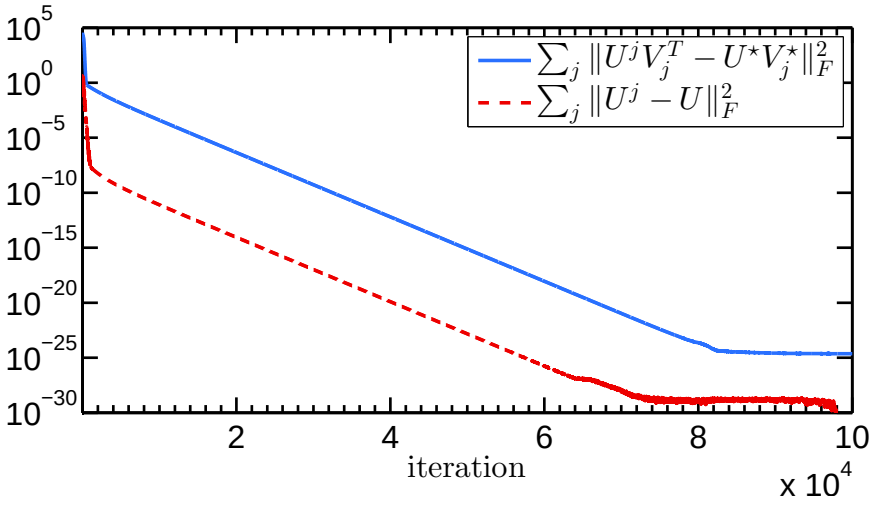


Figure 7.1: Solving (7.17) by using GADMM (7.13).

We apply GADMM to solve (7.17). [199, Theorem 3.1] shows that, under suitable conditions, GADMM is guaranteed to converge to a second-order critical point of (7.17). Although we do not confirm those conditions for the matrix factorization problem (7.17), we use numerical simulations to illustrate the ability of GADMM to reach

solutions that correspond to global minimizers of the centralized problem (7.16).

To set up the experiments, we first generate the rank-r ground truth matrix $\mathbf{Y}^{\#} = [\mathbf{Y}_1^{\#} \ \cdots \ \mathbf{Y}_J^{\#}] \in \mathbb{R}^{n \times Jm}$ ($m = \sum_{j=1}^{J} m_j$) where $r = 2$, $n = 50$, $J = 10$, $m_j = 20 \ \forall j$ by multiplying two standard Gaussian matrices (i.e., each entry i.i.d. from $\mathcal{N}(0,1)$) of size $n \times r$ and $r \times m$, respectively. Then adding a noise matrix $\mathbf{N} \in \mathbb{R}^{n \times m}$ with each entry i.i.d. drawn from $\mathcal{N}(0, \sigma_Z^2)$ with $\sigma_Z = 0.1$, we get the noisy observation $\mathbf{Y} = \mathbf{Y}^{\#} + \mathbf{N}$. In this case, the signal-to-noise ratio can be computed as $\mathrm{SNR} = 10 \log_{10} \left(\mathbb{E} \left[\|\mathbf{Y}^{\#}\|_F^2 \right] / \mathbb{E} \left[\|\mathbf{N}\|_F^2 \right] \right) = 10 \log_{10} (\frac{r}{\sigma_Z^2}) = 23 \, \mathrm{dB}$.

To estimate the ground truth, we then solve (7.17) with $\mu = 1$ by using GADMM (7.13) with $\rho = 10$, $\beta = 1000$ and a random initialization. To verify our main results (cf. Corollary 7.5.1), we plot the optimality distance $\sum_{j=1}^{J} \|\mathbf{U}^j \mathbf{V}_j^\top - \mathbf{U}^\star \mathbf{V}_j^\star\|_F^2$ and consensus error $\sum_{j=1}^{J} \|\mathbf{U}^j - \mathbf{U}\|_F^2$ as a function of the number of iterations, where $(\mathbf{U}^\star, [\mathbf{V}_1^\star \ \cdots \ \mathbf{V}_J^\star])$ is a global minimizer of problem (7.16). Figure 7.1 shows that the GADMM achieves both global optimum and exact consensus.

CHAPTER 8

GLOBAL OPTIMALITY IN DISTRIBUTED LOW-RANK MATRIX FACTORIZATION

We study the convergence of a variant of distributed gradient descent (DGD) on a distributed low-rank matrix approximation problem wherein some optimization variables are used for consensus (as in classical DGD) and some optimization variables appear only locally at a single node in the network. We term the resulting algorithm DGD+LOCAL. Using algorithmic connections to gradient descent and geometric connections to the well-behaved landscape of the centralized low-rank matrix approximation problem, we identify sufficient conditions where DGD+LOCAL is guaranteed to converge with exact consensus to a global minimizer of the original centralized problem. For the distributed low-rank matrix approximation problem, these guarantees are stronger—in terms of consensus and optimality—than what appear in the literature for classical DGD and more general problems.

8.1 Introduction

A promising line of recent literature has examined the nonconvex objective functions that arise when certain matrix optimization problems are solved in factored form, that is, when a low-rank optimization variable $\mathbf{X}$ is replaced by a product of two thin matrices $\mathbf{U}\mathbf{V}^\top$ and the optimization proceeds jointly over $\mathbf{U}$ and $\mathbf{V}$ [6,8,93,101,140,192,204]. In many cases, a study of the geometric landscape of these objective functions reveals that—despite their nonconvexity—they possess a certain favorable geometry. In particular, many of the resulting objective functions (i) satisfy the *strict saddle property* [97, 114], where every critical point is either a local minimum or is a strict saddle point, at which the Hessian matrix has at least one negative eigenvalue, and (ii) have no spurious local minima (every local minimum corresponds to a global minimum).

One such problem—which is both of fundamental importance and representative of structures that arise in many other machine learning problems [205]—is the low-rank matrix approximation problem, where given a data matrix $\mathbf{Y}$ the objective is to minimize $\|\mathbf{U}\mathbf{V}^\top - \mathbf{Y}\|_F^2$. As we explain in Theorem 8.3.1, building on recent analysis in [206] and [93], this problem satisfies the strict saddle property and has no spurious local minima.

In parallel with the recent focus on the favorable geometry of certain nonconvex landscapes, it has been shown that a number of local search algorithms have the capability to avoid strict saddle points and converge to a local minimizer for problems that satisfy the strict saddle property [113, 126, 127, 207]. As stated in [127] and as we summarize in Theorems 8.2.2 and 8.2.4, gradient descent when started from a random initialization is one such algorithm. For problems such as low-rank matrix approximation that have no spurious local minima, converging to a local minimizer means converging to a global minimizer.

To date, the geometric and algorithmic research described above has largely focused on *centralized optimization*, where all computations happen at one "central" node that has full access, for example, to the data matrix $\mathbf{Y}$.

In this work, we study the impact of *distributing* the factored optimization problem, such as would be necessary if the data matrix $\mathbf{Y}$ in low-rank matrix approximation were partitioned into submatrices $\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 & \cdots & \mathbf{Y}_J \end{bmatrix}$, each of which was available at only one node in a network. By similarly partitioning the matrix $\mathbf{V}$, one can partition the objective function

$$\|\mathbf{U}\mathbf{V}^\top - \mathbf{Y}\|_F^2 = \sum_{j=1}^{J} \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2. \tag{8.1}$$

As we discuss, one can attempt to minimize the resulting objective, in which the matrix $\mathbf{U}$ appears in every term of the summation, using techniques similar to classical distributed algorithms such as distributed gradient descent (DGD) [208]. These algorithms, however, involve creating local copies $\mathbf{U}^1, \mathbf{U}^2, \ldots, \mathbf{U}^J$ of the optimization variable $\mathbf{U}$ and iteratively sharing updates of these variables with the aim of converging to a consensus where (exactly or approximately) $\mathbf{U}^1 = \mathbf{U}^2 = \cdots = \mathbf{U}^J$.

In this work we study a straightforward extension of DGD for solving such problems. This extension, which we term DGD+LOCAL, resembles classical DGD in that each node j has a local *copy* $\mathbf{U}^j$ of the optimization variable $\mathbf{U}$ as described above. Additionally, however, each node has a local *block* $\mathbf{V}_j$ of the partitioned optimization variable $\mathbf{V}$, and this block exists only locally at node j without any consensus or sharing among other nodes.

We present a geometric framework for analyzing the convergence of DGD+LOCAL in such problems. Our framework relies on a straightforward conversion which reveals (for example in the low-rank matrix approximation problem) that DGD+LOCAL as described above is equivalent to running conventional gradient descent on the objective function

$$\sum_{j=1}^{J} \left(\|\mathbf{U}^j\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2 + \sum_{i=1}^{J} w_{ji} \|\mathbf{U}^j - \mathbf{U}^i\|_F^2 \right), \tag{8.2}$$

where w_{ji} are weights inherited from the DGD+LOCAL iterations. This objective function (8.2) differs from the original objective function (8.1) in two respects: it contains more optimization variables, and it includes a quadratic regularizer to encourage consensus. Although the geometry of (8.1) is understood to be well-behaved, new questions arise about the geometry of (8.2): Does it contain new critical points (local minima that are not global, saddle points that are not strict)? And on the consensus subspace, where $\mathbf{U}^1 = \mathbf{U}^2 = \cdots = \mathbf{U}^J$, how do the critical points of (8.2) relate to the critical points of (8.1)? We answer these questions and build on the algorithmic results for gradient descent to identify in Theorem 8.3.2 sufficient conditions where DGD+LOCAL is guaranteed to converge to a point that (i) is exactly on the consensus subspace, and (ii) coincides with a global minimizer of problem (8.1). Under these conditions, the distributed low-rank matrix approximation problem is shown to enjoy the same geometric and algorithmic guarantees as its well-behaved centralized counterpart.

For the distributed low-rank matrix approximation problem, these guarantees are stronger than what appear in the literature for classical DGD and more general problems. In particular, we show exact convergence to the consensus subspace with a fixed DGD+LOCAL stepsize, which in more general works is accomplished only with diminishing DGD stepsizes for convex [200, 201] and nonconvex [202] problems or by otherwise modifying DGD as in the EXTRA algorithm [209]. Moreover, we show convergence to a global minimizer of the original centralized nonconvex problem. Until recently, existing DGD results either considered convex problems [200, 201] or showed convergence to stationary points of nonconvex problems [202]. Very recently, it was also shown [203] that with an appropriately small stepsize, DGD can converge to an arbitrarily small neighborhood of a second-order critical point for general nonconvex problems with additional technical assumptions. Our work differs from [203] in our use of DGD+LOCAL (rather than DGD) and our focus on one specific problem where we can establish stronger guarantees of exact global optimality and exact consensus without requiring an arbitrarily small (or diminishing) stepsize.

Our main results on distributed low-rank matrix factorization are presented in Section 8.3. These results build on several more general algorithmic and geometric results that we first establish in Section 8.2. The results from Section 8.2 may have broader applicability, and the geometric and algorithmic discussions in Section 8.2 may have independent interest from one another.

8.2 General Analysis of DGD+LOCAL

Consider a centralized minimization problem that can be written in the form

$$\underset{\mathbf{x},\mathbf{y}}{\text{minimize}}\, f(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{J} f_j(\mathbf{x},\mathbf{y}_j), \tag{8.3}$$

where $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1^\top & \cdots & \mathbf{y}_J^\top \end{bmatrix}^\top$. Here $\mathbf{x}$ is the common variable in all of the objective functions $\{f_j\}_{j\in[J]}$ and $\mathbf{y}_j$ is the variable only corresponding to f_j.

The standard DGD algorithm [208] is stated for problems of the form

$$\underset{\mathbf{x}}{\text{minimize}}\, f(\mathbf{x}) = \sum_{j=1}^{J} f_j(\mathbf{x}),$$

and for such problems it involves updates of the form

$$\mathbf{x}^j(k+1) = \sum_{i=1}^{J} \left(\widetilde{w}_{ji}\mathbf{x}^i(k)\right) - \mu\nabla_{\mathbf{x}} f_j(\mathbf{x}^j(k)),$$

where $\{\widetilde{w}_{ji}\}$ are a set of symmetric nonnegative weights, and $\widetilde{w}_{ji}$ is positive if and only if nodes i and j are neighbors in the network or $i = j$. Throughout this chapter, we will make the common assumption [210] that

$$\sum_{i=1}^{J} \widetilde{w}_{ji} = 1 \text{ for all } j \in [J]. \tag{8.4}$$

A very natural extension of DGD to problems of the form (8.3)—which involve local *copies* of the shared variable $\mathbf{x}$ and local *partitions* of the variable $\mathbf{y}$—is to perform the updates

$$\begin{aligned} \mathbf{x}^j(k+1) &= \sum_{i=1}^{J} \left(\widetilde{w}_{ji}\mathbf{x}^i(k)\right) - \mu \nabla_{\mathbf{x}} f_j(\mathbf{x}^j(k), \mathbf{y}_j(k)), \\ \mathbf{y}_j(k+1) &= \mathbf{y}_j(k) - \mu \nabla_{\mathbf{y}} f_j(\mathbf{x}^j(k), \mathbf{y}_j(k)). \end{aligned} \tag{8.5}$$

Because we are interested in solving problems of the form (8.3), we refer to (8.5) as DGD+LOCAL throughout this chapter. We note that DGD+LOCAL is not equivalent to algorithm would obtain by applying classical DGD to reach consensus over the concatenated variables $\mathbf{x}$ and $\mathbf{y}$ as this would require each node to maintain a local copy of the entire vector $\mathbf{y}$. For the same reason, DGD+LOCAL is not equivalent to the blocked variable problem described in [211].

8.2.1 Relation to Gradient Descent

Note that we can rewrite the first equation in (8.5) as

$$\begin{aligned} \mathbf{x}^j(k+1) &= (\sum_{i=1}^{J} \widetilde{w}_{ji})\mathbf{x}^j(k) - \mu \left(\nabla_{\mathbf{x}} f_j(\mathbf{x}^j(k), \mathbf{y}_j(k)) + \sum_{i \neq j} \frac{\widetilde{w}_{ji}}{\mu}(\mathbf{x}^j(k) - \mathbf{x}^i(k)) \right) \\ &= \mathbf{x}^j(k) - \mu \left(\nabla_{\mathbf{x}} f_j(\mathbf{x}^j(k), \mathbf{y}_j(k)) + \sum_{i \neq j} \frac{\widetilde{w}_{ji}}{\mu}(\mathbf{x}^j(k) - \mathbf{x}^i(k)) \right). \end{aligned}$$

In the second line, we have used the assumption (8.4). Thus, by defining $\{w_{ji}\}$ such that

$$w_{ji} = w_{ij} = \begin{cases} \frac{\widetilde{w}_{ji}}{4\mu}, & i \neq j, \\ 0, & i = j, \end{cases} \tag{8.6}$$

we see that DGD+LOCAL (8.5) is equivalent to applying standard gradient descent (with stepsize μ) to the problem

$$\underset{\mathbf{z}}{\text{minimize}}\, g(\mathbf{z}) = \sum_{j=1}^{J} \left(f_j(\mathbf{x}^j, \mathbf{y}_j) + \sum_{i=1}^{J} w_{ji} \|\mathbf{x}^j - \mathbf{x}^i\|_2^2 \right), \tag{8.7}$$

where $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^J, \mathbf{y}_1, \ldots, \mathbf{y}_J)$ and $\mathbf{W} = \{w_{ji}\}$ is a $J \times J$ connectivity matrix with nonnegative entries defined in (8.6) and zeros on the diagonal.

8.2.2 Algorithmic Analysis

We are interested in understanding the convergence of the gradient descent algorithm when it is applied to minimizing $g(\mathbf{z})$ in (8.7); as we have argued in Section 8.2.1, this is equivalent to running the DGD+LOCAL algorithm (8.5) to minimize the objective function $f(\mathbf{x}, \mathbf{y})$ in (8.3).

Under certain conditions, we can guarantee that gradient descent will converge to a second-order critical point of the objective function $g(\mathbf{z})$ in (8.7). The proof relies on certain properties of the functions f_j comprising (8.3). We first describe these properties before providing the convergence result.

8.2.2.1 Objective Function Properties and Convergence of Gradient Descent

The first property concerns the assumption that each f_j comprising (8.3) has Lipschitz gradient. In this case we can also argue that g in (8.7) has Lipschitz gradient.

Proposition 8.2.1. *Let $f(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{J} f_j(\mathbf{x}, \mathbf{y}_j)$ be an objective function as in* (8.3) *and let $g(\mathbf{z})$ be as in* (8.7) *with $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^J, \mathbf{y}_1, \ldots, \mathbf{y}_J)$. Suppose that each f_j has Lipschitz gradient, i.e., ∇f_j is Lipschitz continuous with constant $L_j > 0$. Then ∇g is Lipschitz continuous with constant*

$$L_g = L + \frac{2\omega}{\mu},$$

where $L := \max_j L_j$, $\omega := \sum_{i \neq j}^{J} \widetilde{w}_{ji}$, and $\widetilde{w}_{ji}$ and μ are the DGD+LOCAL weights and stepsize as in (8.5).

Proposition 8.2.1 is proved in Appendix F.1.

The second property concerns the following Łojasiewicz inequality, which arises in the convergence analysis of gradient descent.

Definition 8.2.1. *[212] Assume that $h : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Then h is said to satisfy the Łojasiewicz inequality, if for any critical point $\overline{\mathbf{x}}$ of $h(\mathbf{x})$, there exist $\delta > 0,\ \theta \in [0, 1),\ C_1 > 0$ such that*

$$|h(\mathbf{x}) - h(\overline{\mathbf{x}})|^{\theta} \leq C_1 \|\nabla h(\mathbf{x})\|, \ \ \forall\, \mathbf{x} \in B(\overline{\mathbf{x}}, \delta).$$

Here θ is often referred to as the KL exponent.

This Łojasiewicz inequality (or a more general Kurdyka-Łojasiewicz (KL) inequality for the general nonsmooth problems) characterizes the local geometric properties of the objective function around its critical points and has proved useful for convergence analysis [212, 213]. The Łojasiewicz inequality (or KL inequality) is very general and holds for most problems in engineering. For example, every analytic function satisfies this Łojasiewicz inequality, but each function may have different Łojasiewicz exponent θ which determines the convergence rate; see [212, 213] for the details on this.

A general result for convergence of gradient descent to first-order critical point for a function satisfying the Łojasiewicz inequality is as follows.[25]

Theorem 8.2.1. *[212] Suppose* $\inf_{\mathbb{R}^n} h > -\infty$ *and* h *satisfies the Łojasiewicz inequality. Also assume* ∇h *is Lipschitz continuous with constant* $L > 0$*. Let* $\{\mathbf{x}(k)\}$ *be the sequence generated by gradient descent* $\mathbf{x}(k+1) = \mathbf{x}(k) - \mu \nabla h(\mathbf{x}(k))$ *with* $\mu < \frac{1}{L}$*. Then if the sequence* $\{\mathbf{x}(k)\}$ *is bounded, it converges to a critical point of* h*.*

The following result further characterizes the convergence behavior of gradient descent to a second-order critical point.

Theorem 8.2.2. *[113] Suppose* h *is a twice-continuously differentiable function and* ∇h *is Lipschitz continuous with constant* $L > 0$*. Let* $\{\mathbf{x}(k)\}$ *be the sequence generated by gradient descent* $\mathbf{x}(k+1) = \mathbf{x}(k) - \mu \nabla h(\mathbf{x}(k))$ *with* $\mu < \frac{1}{L}$*. Suppose* $\mathbf{x}(0)$ *is chosen randomly from a probability distribution supported on a set* S *having positive measure. Then the sequence* $\{\mathbf{x}(k)\}$ *almost surely avoids strict saddles, where the Hessian has at least one negative eigenvalue.*

Theorems 8.2.1 and 8.2.2 apply for functions h that globally satisfy the Łojasiewicz and Lipschitz gradient conditions. In some problems, however, one or both of these properties may be satisfied only locally. Nevertheless, under an assumption of bounded iterations—as is already made in Theorem 8.2.1—it is possible to extend the first- and second-order convergence results to such functions. For example, one can extend Theorem 8.2.1 as follows by noting that the original derivation in [212] used the Łojasiewicz property only locally around limit points of the sequence $\{\mathbf{x}(k)\}$.

Theorem 8.2.3. *[212] Suppose* $\inf_{\mathbb{R}^n} h > -\infty$*. For* $\rho > 0$*, let* B_ρ *denote the open ball of radius* ρ*:*

$$B_\rho := \{\mathbf{x} : \|\mathbf{x}\|_2 < \rho\},$$

and suppose h *satisfies the Łojasiewicz inequality at all points* $\mathbf{x} \in B_\rho$*. Also assume* ∇h *is Lipschitz continuous with constant* $L > 0$*. Let* $\{\mathbf{x}(k)\}$ *be the sequence generated by gradient descent* $\mathbf{x}(k+1) = \mathbf{x}(k) - \mu \nabla h(\mathbf{x}(k))$ *with* $\mu < \frac{1}{L}$*. Suppose* $\{\mathbf{x}(k)\} \subseteq B_\rho$ *and all limit points of* $\{\mathbf{x}(k)\}$ *are in* B_ρ*. Then the sequence* $\{\mathbf{x}(k)\}$ *converges to a critical point of* h*.*

The following result establishes second-order convergence for a function with a locally Lipschitz gradient.

Theorem 8.2.4. *Let* $\rho > 0$*, and consider an objective function* h *where:*

1. $\inf_{\mathbb{R}^n} h > -\infty$*,*
2. h *satisfies the Łojasiewicz inequality within* B_ρ*,*

[25] The result in [212] is stated for the proximal method, but the result can be extended to gradient descent as long as $\mu < \frac{1}{L}$.

3. h is twice-continuously differentiable, and

4. $|h(\mathbf{x})| \leq L_0$, $\|\nabla h(\mathbf{x})\| \leq L_1$, and $\left\|\nabla^2 h(\mathbf{x})\right\|_2 \leq L_2$ for all $\mathbf{x} \in B_{2\rho}$.

Suppose the gradient descent stepsize

$$\mu < \frac{1}{L_2 + \frac{4L_1}{\rho} + \frac{(4+2\pi)L_0}{\rho^2}}. \tag{8.8}$$

Suppose $\mathbf{x}(0)$ is chosen randomly from a probability distribution supported on a set $S \subseteq B_\rho$ with S having positive measure, and suppose that under such random initialization, there is a positive probability that the sequence $\{\mathbf{x}(k)\}$ remains bounded in B_ρ and all limit points of $\{\mathbf{x}(k)\}$ are in B_ρ.

Then conditioned on observing that $\{\mathbf{x}(k)\} \subseteq B_\rho$ and all limit points of $\{\mathbf{x}(k)\}$ are in B_ρ, gradient descent converges to a critical point of h, and the probability that this critical point is a strict saddle point is zero.

Theorem 8.2.4 is proved in Appendix F.2.

8.2.2.2 Convergence Analysis of DGD+LOCAL

As described in the following theorem, under certain conditions, we can guarantee that the DGD+LOCAL algorithm (8.5) (which is equivalent to gradient descent applied to minimizing $g(\mathbf{z})$ in (8.7)) will converge to a second-order critical point of the objective function $g(\mathbf{z})$.

Theorem 8.2.5. *Let $f(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{J} f_j(\mathbf{x}, \mathbf{y}_j)$ be an objective function as in* (8.3) *and let $g(\mathbf{z})$ be as in* (8.7) *with $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^J, \mathbf{y}_1, \ldots, \mathbf{y}_J)$. Suppose each f_j satisfies $\inf_{\mathbb{R}^n} f_j > -\infty$, is twice continuously-differentiable, and has Lipschitz gradient, i.e., ∇f_j is Lipschitz continuous with constant $L_j > 0$. Suppose g satisfies the Łojasiewicz inequality. Let $L := \max_j L_j$, and let $\widetilde{w}_{ji}$ and μ be the DGD+LOCAL weights and stepsize as in* (8.5)*.*

Assume $\omega := \max_j \sum_{i \neq j} \widetilde{w}_{ji} < \frac{1}{2}$. Let $\{\mathbf{z}(k)\}$ be the sequence generated by the DGD+LOCAL algorithm in (8.5) *with*

$$\mu < \frac{1 - 2\omega}{L} \tag{8.9}$$

and with random initialization from a probability distribution supported on a set S having positive measure. Then if the sequence $\{\mathbf{z}(k)\}$ is bounded, it almost surely converges to a second-order critical point of the objective function in (8.7)*.*

Proof. Recall that running the DGD+LOCAL algorithm (8.5) to minimize the objective function $f(\mathbf{x}, \mathbf{y})$ in (8.3) is equivalent to running gradient descent on $g(\mathbf{z})$ in (8.7). The proof is completed by invoking Theorem 8.2.1 and Theorem 8.2.2 with h replaced by g. From Proposition 8.2.1, we have that ∇g is Lipschitz continuous with constant

$L_g = L + \frac{2\omega}{\mu}$, and so choosing μ to satisfy (8.9) ensures that $\mu < \frac{1}{L_g}$ as required in Theorem 8.2.1 and Theorem 8.2.2. □

Remark 8.2.1. The requirement that the DGD+LOCAL stepsize $\mu = O(\frac{1}{L})$ also appears in the convergence analysis of DGD in [202, 214].

Remark 8.2.2. The function g is guaranteed to satisfy the Łojasiewicz inequality, for example, if every f_j is semi-algebraic, because this will imply that g is semi-algebraic, and every semi-algebraic function satisfies the Łojasiewicz inequality.

Remark 8.2.3. In order to satisfy (8.9), it must hold that $\omega < \frac{1}{2}$. In the case where the DGD+LOCAL weight matrix $\widetilde{\mathbf{W}}$ is symmetric and doubly stochastic (i.e., $\widetilde{\mathbf{W}}$ has nonnegative entries and each of its rows and columns sums to 1), this condition is equivalent to requiring that each diagonal element of $\widetilde{\mathbf{W}}$ is larger than $\frac{1}{2}$. Given any symmetric and doubly stochastic matrix $\widetilde{\mathbf{W}}$, one can design a new weight matrix $(\widetilde{\mathbf{W}} + \mathbf{I})/2$ that satisfies this requirement. This strategy is also mentioned at the end of [214, Section 2.1].

We also have the following DGD+LOCAL convergence result when the functions f_j have only a locally Lipschitz gradient.

Theorem 8.2.6. *Let $f(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{J} f_j(\mathbf{x}, \mathbf{y}_j)$ be an objective function as in* (8.3) *and let $g(\mathbf{z})$ be as in* (8.7) *with $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^J, \mathbf{y}_1, \ldots, \mathbf{y}_J)$. Let $\rho > 0$ and suppose each f_j satisfies*

1. $\inf_{\mathbb{R}^n} f_j > -\infty$,
2. *f_j is twice-continuously differentiable, and*
3. $|f_j(\mathbf{x}, \mathbf{y}_j)| \leq L_{0,j}$, $\|\nabla f_j(\mathbf{x}, \mathbf{y}_j)\| \leq L_{1,j}$, *and* $\left\|\nabla^2 f_j(\mathbf{x}, \mathbf{y})\right\|_2 \leq L_{2,j}$ *for all* $(\mathbf{x}, \mathbf{y}_j) \in B_{2\rho}$.

Suppose also that g satisfies the Łojasiewicz inequality within B_ρ. Let $\widetilde{w}_{ji}$ and μ be the DGD+LOCAL weights and stepsize as in (8.5). *Assume $\omega := \max_j \sum_{i \neq j} \widetilde{w}_{ji} < \frac{1}{2}$. Let $\{\mathbf{z}(k)\}$ be the sequence generated by the DGD+LOCAL algorithm in* (8.5) *with*

$$\mu < \frac{1 - 2\omega}{\max_j L_{2,j} + \frac{4L_{1,j}}{\rho} + \frac{(4+2\pi)L_{0,j}}{\rho^2}}. \tag{8.10}$$

Suppose $\mathbf{z}(0)$ is chosen randomly from a probability distribution supported on a set $S \subseteq B_\rho$ with S having positive measure, and suppose that under such random initialization, there is a positive probability that the sequence $\{\mathbf{z}(k)\}$ remains bounded in B_ρ and all limit points of $\{\mathbf{z}(k)\}$ are in B_ρ.

Then conditioned on observing that $\{\mathbf{z}(k)\} \subseteq B_\rho$ and all limit points of $\{\mathbf{z}(k)\}$ are in B_ρ, DGD+LOCAL converges to a critical point of the objective function in (8.7)*, and the probability that this critical point is a strict saddle point is zero.*

Proof. Recall that running the DGD+LOCAL algorithm (8.5) to minimize the objective function $f(\mathbf{x}, \mathbf{y})$ in (8.3) is equivalent to running gradient descent on $g(\mathbf{z})$ in (8.7). Similar to the approach taken in proving Theorem 8.2.4, to deal with the local Lipschitz condition, the proof involves constructing a function $\widetilde{g}$ such that $\widetilde{g}(\mathbf{z}) = g(\mathbf{z})$ for all $\mathbf{z} \in B_\rho$ but where $\widetilde{g}$ has a globally Lipschitz gradient.

To do this, recall the window function w defined in Appendix F.2. Now, recall that

$$g(\mathbf{z}) = \sum_{j=1}^{J} \left(f_j(\mathbf{x}^j, \mathbf{y}_j) + \sum_{i=1}^{J} w_{ji} \|\mathbf{x}^j - \mathbf{x}^i\|_2^2 \right)$$

and define

$$\widetilde{g}\,(\mathbf{z}) = \sum_{j=1}^{J} \left(\widetilde{f}_j(\mathbf{x}^j, \mathbf{y}_j) + \sum_{i=1}^{J} w_{ji} \|\mathbf{x}^j - \mathbf{x}^i\|_2^2 \right), \tag{8.11}$$

where

$$\widetilde{f}_j(\mathbf{x}^j, \mathbf{y}_j) = f_j(\mathbf{x}^j, \mathbf{y}_j) w(\left[(\mathbf{x}^j)^\top \quad \mathbf{y}_j^\top \right]^\top).$$

Since $\widetilde{f}_j(\mathbf{x}^j, \mathbf{y}_j) = f_j(\mathbf{x}^j, \mathbf{y}_j)$ for $(\mathbf{x}^j, \mathbf{y}_j) \in B_\rho$, we have that $\widetilde{g}\,(\mathbf{z}) = g(\mathbf{z})$ for all $\mathbf{z} \in B_\rho$.

We have the following properties for $\widetilde{g}$:

- Since $g = \widetilde{g}$ in B_ρ, $\widetilde{g}$ satisfies the Łojasiewicz inequality in B_ρ.
- Since $f_j \in C^2$ for all j and $w \in C^2$, $\widetilde{g} \in C^2$.
- Since $\inf_{\mathbb{R}^n} f_j > -\infty$ for all j and $\inf_{\mathbb{R}^n} w > -\infty$, $\inf_{\mathbb{R}^n} \widetilde{g} > -\infty$.
- To globally bound the Lipschitz constant of the gradient of $\widetilde{g}$, note that

$$\begin{aligned} \left\| \nabla^2 \widetilde{f}_j \right\| &= \left\| w \cdot \nabla^2 f_j + \nabla f_j \cdot (\nabla w)^\top + \nabla w \cdot (\nabla f_j)^\top + f_j \cdot \nabla^2 w \right\| \\ &\leq |w| \left\| \nabla^2 f_j \right\| + 2 \|\nabla w\| \|\nabla f_j\| + |f_j| \left\| \nabla^2 w \right\| \\ &\leq L_{2,j} + \frac{4L_{1,j}}{\rho} + \frac{(4 + 2\pi) L_{0,j}}{\rho^2} \quad \text{for all } (\mathbf{x}^j, \mathbf{y}_j). \end{aligned}$$

Therefore, given the form of $\widetilde{g}$ in (8.11), we can conclude from Proposition 8.2.1 that globally, $\nabla \widetilde{g}$ is Lipschitz continuous with constant

$$L_{\widetilde{g}} = \left(\max_j L_{2,j} + \frac{4L_{1,j}}{\rho} + \frac{(4 + 2\pi) L_{0,j}}{\rho^2} \right) + \frac{2\omega}{\mu}.$$

Now consider the gradient descent algorithm with stepsize μ satisfying (8.10). Define

$$\begin{aligned} T_g = \{ \mathbf{z}(0) \in B_\rho : &\text{ all } \{\mathbf{z}(k)\} \subseteq B_\rho \text{ and all limit points of } \{\mathbf{z}(k)\} \text{ are in } B_\rho \\ &\text{when gradient descent is run on } g \text{ starting at } \mathbf{z}(0) \} \end{aligned}$$

and

$$T_{\widetilde{g}} = \{\mathbf{z}(0) \in B_\rho : \text{all } \{\mathbf{z}(k)\} \subseteq B_\rho \text{ and all limit points of } \{\mathbf{z}(k)\} \text{ are in } B_\rho \\ \text{when gradient descent is run on } \widetilde{g} \text{ starting at } \mathbf{z}(0)\}.$$

Similarly, define

$$\Sigma_g = \{\mathbf{z}(0) \in B_\rho : \ \{\mathbf{z}(k)\} \text{ converges to a strict saddle when gradient descent is run on } g \text{ starting at } \mathbf{z}(0)\}$$

and

$$\Sigma_{\widetilde{g}} = \{\mathbf{z}(0) \in B_\rho : \ \{\mathbf{z}(k)\} \text{ converges to a strict saddle when gradient descent is run on } \widetilde{g} \text{ starting at } \mathbf{z}(0)\}.$$

Using the above properties, we see that Theorem 8.2.2 can be applied to $\widetilde{g}$, and so we conclude that $\Sigma_{\widetilde{g}}$ has measure zero.

Now, after running gradient descent on g from a random initialization as in the theorem statement, condition on observing that $\{\mathbf{z}(k)\} \subseteq B_\rho$ and all limit points of $\{\mathbf{z}(k)\}$ are in B_ρ, i.e., that $\mathbf{z}(0) \in T_g$. Because $\{\mathbf{z}(k)\} \subseteq B_\rho$ and all limit points of $\{\mathbf{z}(k)\}$ are in B_ρ, and because $\{\mathbf{z}(k)\}$ matches the sequence that would be obtained by running gradient descent on $\widetilde{g}$, we can apply Theorem 8.2.3 to conclude that $\{\mathbf{z}(k)\}$ converges to a critical point of $\widetilde{g}$, and since this critical point belongs to B_ρ and $\widetilde{g} = g$ inside B_ρ, we conclude that this is also a critical point of g.

Finally, using the definition of conditional probability, we have

$$\begin{aligned} P(\mathbf{z}(0) \in \Sigma_g | \mathbf{z}(0) \in T_g) &= \frac{P(\mathbf{z}(0) \in \Sigma_g \cap T_g)}{P(\mathbf{z}(0) \in T_g)} \\ &= \frac{P(\mathbf{z}(0) \in \Sigma_{\widetilde{g}} \cap T_{\widetilde{g}})}{P(\mathbf{z}(0) \in T_g)}, \end{aligned}$$

where the second equality follows from the fact that $\widetilde{g} = g$ inside B_ρ: if a sequence of iterations stays bounded inside B_ρ and converges to a strict saddle when gradient descent is run on g, the same will hold when gradient descent is run on $\widetilde{g}$, and vice versa. Since $\Sigma_{\widetilde{g}}$ has zero measure and because $\mathbf{z}(0)$ is chosen randomly from a probability distribution supported on a set $S \subseteq B_\rho$ with S having positive measure, $P(\mathbf{z}(0) \in \Sigma_{\widetilde{g}} \cap T_{\widetilde{g}}) = 0$. Also, by assumption, $P(\mathbf{z}(0) \in T_g) > 0$. Therefore, $P(\mathbf{z}(0) \in \Sigma_g | \mathbf{z}(0) \in T_g) = \frac{0}{\text{nonzero}} = 0$. □

8.2.3 Geometric Analysis

Section 8.2.2 establishes that, under certain conditions, DGD+LOCAL will converge to a second-order critical point of the objective function $g(\mathbf{z})$ in (8.7).

In this section, we are interested in studying the geometric landscape of the distributed objective function in (8.7) and comparing it to the geometric landscape of the original centralized objective function in (8.3). In particular, we

would like to understand how the critical points of $g(\mathbf{z})$ in (8.7)) are related to the critical points of $f(\mathbf{x}, \mathbf{y})$ in (8.3).

These problems differ in two important respects:

- The objective function in (8.7) involves more optimization variables than that in (8.3). Thus, the optimization takes place in a higher-dimensional space and there is the potential for new features to be introduced into the geometric landscape.

- The objective function in (8.7) involves a quadratic regularization term that will promote consensus among the variables $\mathbf{x}^1, \ldots, \mathbf{x}^J$. This term is absent from (8.3). However, along the *consensus subspace* where $\mathbf{x}^1 = \cdots = \mathbf{x}^J$, this regularizer will be zero and the objective functions will coincide.

Despite these differences, we characterize below some ways in which the geometric landscapes of the two problems may be viewed as equivalent. These results may have independent interest from the specific DGD+LOCAL convergence analysis in Section 8.2.2.

Our first result establishes that if the sub-objective functions f_j satisfy certain properties, the formulation (8.7) does not introduce any new global minima outside of the consensus subspace.

Proposition 8.2.2. *Let $f(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{J} f_j(\mathbf{x}, \mathbf{y}_j)$ be as in* (8.3)*. Suppose the topology defined by $\mathbf{W}$ is connected. Also suppose there exist $\mathbf{x}^\star$ (which is independent of j) and $\mathbf{y}_j^\star, j \in [J]$ such that*

$$(\mathbf{x}^\star, \mathbf{y}_j^\star) \in \arg\min_{\mathbf{x}, \mathbf{y}_j} f_j(\mathbf{x}, \mathbf{y}_j), \ \forall\, j \in [J]. \tag{8.12}$$

Then $g(\mathbf{z})$ defined in (8.7) *satisfies*

$$\min_{\mathbf{z}} g(\mathbf{z}) = \min_{\mathbf{x}, \mathbf{y}} f(\mathbf{x}, \mathbf{y}),$$

and $g(\mathbf{z})$ achieves its global minimum only for $\mathbf{z}$ with $\mathbf{x}^1 = \cdots = \mathbf{x}^J$.

Proposition 8.2.2 is proved in Appendix F.3. We note that the assumption in Proposition 8.2.2 is fairly strong, and while there are problems where it can hold, there are also many problems where it will not hold.

Proposition 8.2.2 establishes that, in certain cases, there will exist no global minimizers of the distributed objective function $g(\mathbf{z})$ that fall outside of the consensus subspace. (Moreover, and also importantly, there will *exist* a global minimizer *on* the consensus subspace.) Also relevant is the question of whether there may exist any *other* types of critical points (such as local minima or saddle points) outside of the consensus subspace. Under certain conditions, the following proposition ensures that the answer is no.

Proposition 8.2.3. *Let $f(\mathbf{x}, \mathbf{y})$ be as in* (8.3) *and $g(\mathbf{z})$ be as in* (8.7) *with $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^J, \mathbf{y}_1, \ldots, \mathbf{y}_J)$. Suppose the matrix $\mathbf{W}$ is connected and symmetric. Also suppose the gradient of f_j satisfies the following symmetric property:*

$$\langle \nabla_{\mathbf{x}} f_j(\mathbf{x}, \mathbf{y}_j), \mathbf{x} \rangle = \langle \nabla_{\mathbf{y}_j} f_j(\mathbf{x}, \mathbf{y}_j), \mathbf{y}_j \rangle \tag{8.13}$$

for all $j \in [J]$. Then, any critical point of g must satisfy $\mathbf{x}^1 = \cdots = \mathbf{x}^J$.

Proposition 8.2.3 is proved in Appendix F.4.

Finally, we can also make a statement about the behavior of critical points that do fall on the consensus subspace.

Theorem 8.2.7. *Let $\mathcal{C}_f$ denote the set of critical points of* (8.3)*:*

$$\mathcal{C}_f := \{\mathbf{x}, \mathbf{y} : \nabla f(\mathbf{x}, \mathbf{y}) = \mathbf{0}\},$$

and let $\mathcal{C}_g$ denote the set of critical points of (8.7)*:*

$$\mathcal{C}_g := \left\{\mathbf{z} : \nabla g(\mathbf{z}) = \mathbf{0}\right\}.$$

Then, for any $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^J, \mathbf{y}) \in \mathcal{C}_g$ with $\mathbf{x}^1 = \cdots = \mathbf{x}^J = \mathbf{x}$, we have $(\mathbf{x}, \mathbf{y}) \in \mathcal{C}_f$. Furthermore, if $(\mathbf{x}, \mathbf{y})$ is a strict saddle of f, then $\mathbf{z} = (\mathbf{x}, \ldots, \mathbf{x}, \mathbf{y})$ is also a strict saddle of g.

The proof of Theorem 8.2.7 is in Appendix F.5.

8.3 Analysis of Distributed Matrix Factorization

We now consider the prototypical low-rank matrix approximation in factored form, where given a data matrix $\mathbf{Y} \in \mathbb{R}^{n \times m}$, we seek to solve

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{V} \in \mathbb{R}^{m \times r}}{\text{minimize}} \ \|\mathbf{U}\mathbf{V}^\top - \mathbf{Y}\|_F^2. \tag{8.14}$$

Here $\mathbf{U} \in \mathbb{R}^{n \times r}$ and $\mathbf{V} \in \mathbb{R}^{m \times r}$ are tall matrices, and r is chosen in advance to allow for a suitable approximation of $\mathbf{Y}$. In some of our results below, we will assume that the data matrix $\mathbf{Y}$ has rank at most r.

One can solve problem (8.14) using local search algorithms such as gradient descent. Such algorithms do not require expensive SVDs, and the storage complexity for $\mathbf{U}$ and $\mathbf{V}$ scales with $(n + m)r$, which is smaller than nm as for $\mathbf{Y}$. Unfortunately, problem (8.14) is nonconvex in the optimization variables $(\mathbf{U}, \mathbf{V})$. Thus, the question arises of whether local search algorithms such as gradient descent actually converge to a global minimizer of (8.14). Using geometric analysis of the critical points of problem (8.14), however, it is possible to prove convergence to a global minimizer.

In Appendix F.6, building on analysis in [206], we prove the following result about the favorable geometry of the nonconvex problem (8.14).

Theorem 8.3.1. *For any data matrix* $\mathbf{Y}$*, every critical point (i.e., every point where the gradient is zero) of problem* (8.14) *is either a global minimum or a strict saddle point, where the Hessian has at least one negative eigenvalue.*

Such favorable geometry has been used in the literature to show that local search algorithms (particularly gradient descent with random initialization [113]) will converge to a global minimum of the objective function.

8.3.1 Distributed Problem Formulation

We are interested in generalizing the matrix approximation problem from centralized to distributed scenarios. To be specific, suppose the columns of the data matrix $\mathbf{Y}$ are distributed among J nodes/sensors. Without loss of generality, partition the columns of $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 & \cdots & \mathbf{Y}_J \end{bmatrix},$$

where for $j \in \{1, 2, \ldots, J\}$, matrix $\mathbf{Y}_j$ (which is stored at node j) has size $n \times m_j$, and where $m = \sum_{j=1}^{J} m_j$. Partitioning $\mathbf{V}$ similarly as

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1^\top & \cdots & \mathbf{V}_J^\top \end{bmatrix}^\top, \tag{8.15}$$

where $\mathbf{V}_j$ has size $m_j \times r$, we obtain the following optimization problem

$$\underset{\mathbf{U}, \mathbf{V}_1, \ldots, \mathbf{V}_J}{\text{minimize}} \sum_{j=1}^{J} \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2, \tag{8.16}$$

which is exactly equivalent to (8.14). Problem (8.16), in turn, can be written in the form of problem (8.3) by taking

$$\mathbf{x} = \text{vec}(\mathbf{U}), \;\; \mathbf{y}_j = \text{vec}(\mathbf{V}_j), \;\; \text{and } f_j(\mathbf{x}, \mathbf{y}_j) = \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2. \tag{8.17}$$

Consequently, we can use the analysis from Section 8.2 to study the performance of DGD+LOCAL (8.5) when applied to problem (8.16).

For convenience, we note that in this context the DGD+LOCAL iterations (8.5) take the form

$$\begin{aligned} \mathbf{U}^j(k+1) &= \sum_{i=1}^{J} \left(\widetilde{w}_{ji}\mathbf{U}^i(k)\right) - 2\mu(\mathbf{U}^j(k)\mathbf{V}_j^\top(k) - \mathbf{Y}_j)\mathbf{V}_j(k), \\ \mathbf{V}_j(k+1) &= \mathbf{V}_j(k) - 2\mu(\mathbf{U}^j(k)\mathbf{V}_j^\top(k) - \mathbf{Y}_j)^\top \mathbf{U}^j(k), \end{aligned} \tag{8.18}$$

and the corresponding gradient descent objective function (8.7) takes the form

$$\underset{\mathbf{z}}{\text{minimize}}\, g(\mathbf{z}) = \sum_{j=1}^{J} \left(\|\mathbf{U}^j \mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2 + \sum_{i=1}^{J} w_{ji} \|\mathbf{U}^j - \mathbf{U}^i\|_F^2 \right), \tag{8.19}$$

where $\mathbf{U}^1, \ldots, \mathbf{U}^J \in \mathbb{R}^{n \times r}$ are local copies of the optimization variable $\mathbf{U}$; $\mathbf{V}_1, \ldots, \mathbf{V}_J$ are a partition of $\mathbf{V}$ as in (8.15); and the weights $\{w_{ji}\}$ are determined by $\{\widetilde{w}_{ji}\}$ and μ as in (8.6).

Problems (8.16) and (8.19) (as special cases of problems (8.3) and (8.7), respectively) satisfy many of the assumptions required for the geometric and algorithmic analysis in Section 8.2. We use these facts in proving our main result for the convergence of DGD+LOCAL on the matrix factorization problem.

Theorem 8.3.2. *Suppose rank*$(\mathbf{Y}) \leq r$*. Suppose DGD+LOCAL* (8.18) *is used to solve problem* (8.16)*, with weights* $\{\widetilde{w}_{ji}\}$ *and stepsize*

$$\mu < \frac{1 - 2\omega}{\max_j \, (276 + 64\pi)\rho^2 + 34\|\mathbf{Y}_j\|_F + \frac{(8+4\pi)}{\rho^2}\|\mathbf{Y}_j\|_F^2} \tag{8.20}$$

for some $\rho > 0$ *and where* $\omega := \max_j \sum_{i \neq j} \widetilde{w}_{ji} < \frac{1}{2}$*. Suppose the* $J \times J$ *connectivity matrix* $\mathbf{W} = \{w_{ji}\}$ *(with* w_{ji} *defined in* (8.6)*) is connected and symmetric. Let* $\{\mathbf{z}(k)\}$ *be the sequence generated by the DGD+LOCAL algorithm. Suppose* $\mathbf{z}(0)$ *is chosen randomly from a probability distribution supported on a set* $S \subseteq B_\rho$ *with* S *having positive measure, and suppose that under such random initialization, there is a positive probability that the sequence* $\{\mathbf{z}(k)\}$ *remains bounded in* B_ρ *and all limit points of* $\{\mathbf{z}(k)\}$ *are in* B_ρ*.*

Then conditioned on observing that $\{\mathbf{z}(k)\} \subseteq B_\rho$ *and all limit points of* $\{\mathbf{z}(k)\}$ *are in* B_ρ*, DGD+LOCAL almost surely converges to a solution* $\mathbf{z}^\star = (\mathbf{U}^{1\star}, \ldots, \mathbf{U}^{J\star}, \mathbf{V}_1^\star, \ldots, \mathbf{V}_J^\star)$ *with the following properties:*

- *Consensus:* $\mathbf{U}^{1\star} = \cdots = \mathbf{U}^{J\star} = \mathbf{U}^\star$*.*
- *Global optimality:* $(\mathbf{U}^\star, \mathbf{V}^\star)$ *is a global minimizer of* (8.14)*, where* $\mathbf{V}^\star$ *denotes the concatenation of* $\mathbf{V}_1^\star, \ldots, \mathbf{V}_J^\star$ *as in* (8.15)*.*

Proof. We begin by arguing that DGD+LOCAL converges almost surely (when $\mathbf{z}(0)$ is chosen randomly inside B_ρ) to a second-order critical point of (8.19). To do this, our goal is to invoke Theorem 8.2.6. We note that each f_j defined in (8.17) satisfies $\inf_{\mathbf{U},\mathbf{V}_j} f_j > -\infty$ and is twice-continuously differentiable. Also, since the functions f_j are semi-algebraic, g satisfies the Łojasiewicz inequality globally. The functions f_j do not have globally Lipschitz gradient. However, we can find quantities $L_{0,j}$, $L_{1,j}$, $L_{2,j}$ such that $|f_j(\mathbf{x}, \mathbf{y}_j)| \leq L_{0,j}$, $\|\nabla f_j(\mathbf{x}, \mathbf{y}_j)\| \leq L_{1,j}$, and $\left\|\nabla^2 f_j(\mathbf{x}, \mathbf{y})\right\|_2 \leq L_{2,j}$ for all $(\mathbf{x}, \mathbf{y}_j) \in B_{2\rho}$. For $L_{0,j}$:

$$
\begin{aligned}
|f_j(\mathbf{x}, \mathbf{y}_j)| &= \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2 \\
&\leq (\|\mathbf{U}\mathbf{V}_j^\top\|_F + \|\mathbf{Y}_j\|_F)^2 \\
&\leq (\|\mathbf{U}\|_F \|\mathbf{V}^j\|_F + \|\mathbf{Y}_j\|_F)^2 \\
&\leq (4\rho^2 + \|\mathbf{Y}_j\|_F)^2 \\
&\leq 32\rho^4 + 2\|\mathbf{Y}_j\|_F^2.
\end{aligned}
$$

For $L_{1,j}$:

$$
\begin{aligned}
\|\nabla f_j(\mathbf{x}, \mathbf{y}_j)\| &= \left\| \begin{bmatrix} \nabla_{\mathbf{U}} \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2 \\ \nabla_{\mathbf{V}_j} \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2 \end{bmatrix} \right\|_F \\
&= \left\| \begin{bmatrix} 2(\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j)\mathbf{V}_j \\ 2(\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j)^\top \mathbf{U} \end{bmatrix} \right\|_F \\
&\leq 2\left(\|\mathbf{U}\mathbf{V}_j^\top \mathbf{V}_j\|_F + \|\mathbf{Y}_j\mathbf{V}_j\|_F + \|\mathbf{V}_j\mathbf{U}^\top\mathbf{U}\|_F + \|\mathbf{Y}_j^\top\mathbf{U}\|_F\right) \\
&\leq 2\left(8\rho^3 + 2\rho\|\mathbf{Y}_j\|_F + 8\rho^3 + 2\rho\|\mathbf{Y}_j\|_F\right) \\
&= 32\rho^3 + 8\rho\|\mathbf{Y}_j\|_F.
\end{aligned}
$$

For $L_{2,j}$, we can bound the Lipschitz constant of ∇f_j in $B_{2\rho}$ as follows. Denote $\mathbf{D} = \begin{bmatrix} \mathbf{D}_{\mathbf{U}} \\ \mathbf{D}_{\mathbf{V}_j} \end{bmatrix}$. Then

$$
\begin{aligned}
&\frac{1}{2}\|\nabla^2 f_j(\mathbf{U}, \mathbf{V}_j)\| = \frac{1}{2} \max_{\|\mathbf{D}\|_F=1} [\nabla^2 f_j(\mathbf{U}, \mathbf{V}_j)](\mathbf{D}, \mathbf{D}) \\
&= \max_{\|\mathbf{D}\|_F=1} \|\mathbf{D}_{\mathbf{U}}\mathbf{V}_j^\top + \mathbf{U}\mathbf{D}_{\mathbf{V}_j}^\top\|_F^2 + 2\langle \mathbf{U}\mathbf{V}_j^\top, \mathbf{D}_{\mathbf{U}}\mathbf{D}_{\mathbf{V}_j}^\top \rangle - 2\langle \mathbf{Y}_j, \mathbf{D}_{\mathbf{U}}\mathbf{D}_{\mathbf{V}_j}^\top \rangle \\
&\leq \max_{\|\mathbf{D}\|_F=1} \frac{5}{2}(\|\mathbf{V}_j\|_F^2 + \|\mathbf{U}\|_F^2)(\|\mathbf{D}_{\mathbf{U}}\|_F^2 + \|\mathbf{D}_{\mathbf{V}_j}\|_F^2) + \|\mathbf{Y}_j\|_F(\|\mathbf{D}_{\mathbf{U}}\|_F^2 + \|\mathbf{D}_{\mathbf{V}_j}\|_F^2) \\
&\leq \max_{\|\mathbf{D}\|_F=1} (10\rho^2 + \|\mathbf{Y}_j\|_F)(\|\mathbf{D}_{\mathbf{U}}\|_F^2 + \|\mathbf{D}_{\mathbf{V}_j}\|_F^2) = 10\rho^2 + \|\mathbf{Y}_j\|_F,
\end{aligned}
$$

where the last inequality holds because $\|\mathbf{U}\|_F^2 + \|\mathbf{V}_j\|_F^2 \leq 4\rho^2$. Therefore we can bound the Lipschitz constant of ∇f_j as $L_j \leq 20\rho^2 + 2\|\mathbf{Y}_j\|_F$ for all $(\mathbf{U}, \mathbf{V}_j)$ such that $\|\mathbf{U}\|_F^2 + \|\mathbf{V}_j\|_F^2 \leq 4\rho^2$. Now,

$$
\begin{aligned}
L_{2,j} + \frac{4L_{1,j}}{\rho} + \frac{(4+2\pi)\,L_{0,j}}{\rho^2} &= 20\rho^2 + 2\|\mathbf{Y}_j\|_F + \frac{4}{\rho}(32\rho^3 + 8\rho\|\mathbf{Y}_j\|_F) + \frac{(4+2\pi)}{\rho^2}(32\rho^4 + 2\|\mathbf{Y}_j\|_F^2) \\
&= 20\rho^2 + 2\|\mathbf{Y}_j\|_F + 128\rho^2 + 32\|\mathbf{Y}_j\|_F + (128 + 64\pi)\rho^2 + \frac{(8+4\pi)}{\rho^2}\|\mathbf{Y}_j\|_F^2 \\
&= (276 + 64\pi)\rho^2 + 34\|\mathbf{Y}_j\|_F + \frac{(8+4\pi)}{\rho^2}\|\mathbf{Y}_j\|_F^2.
\end{aligned}
$$

Thus, choosing μ to satisfy (8.20) ensures that (8.10) is met.

From Theorem 8.2.6, we then conclude that conditioned on observing that $\{\mathbf{z}(k)\} \subseteq B_\rho$ and all limit points of $\{\mathbf{z}(k)\}$ are in B_ρ, DGD+LOCAL converges to a critical point of the objective function in (8.19), and the probability that this critical point is a strict saddle point is zero. We refer to this point as $\mathbf{z}^\star$.

Next, note that the assumption of Proposition 8.2.2 is satisfied if $\mathbf{Y}$ has rank at most r. In particular, there exist $\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}}$ such that $\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top = \mathbf{Y}$ and so we may take $\mathbf{x}^\star = \mathrm{vec}(\widetilde{\mathbf{U}})$ and $\mathbf{y}_j^\star = \mathrm{vec}(\widetilde{\mathbf{V}}_j)$ to achieve $f_j(\mathbf{x}^\star, \mathbf{y}_j^\star) = 0$, which is the smallest possible value for each f_j. Proposition 8.2.2 thus guarantees that (8.19) has at least one critical point that is not a strict saddle (and in fact that it is a global minimizer that falls on the consensus subspace).

Next, note that the symmetric property required for Proposition 8.2.3 is satisfied. To see this, observe that

$$\nabla_{\mathbf{U}} \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2 = 2(\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j)\mathbf{V}_j$$

and

$$\nabla_{\mathbf{V}_j} \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2 = 2(\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j)^\top \mathbf{U}.$$

Thus,

$$\langle \nabla_{\mathbf{U}} \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2, \mathbf{U} \rangle = 2 \cdot \mathrm{tr}(U^\top (\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j)\mathbf{V}_j) = 2 \cdot \mathrm{tr}(\mathbf{V}_j^\top (\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j)^\top \mathbf{U}) = \langle \nabla_{\mathbf{V}_j} \|\mathbf{U}\mathbf{V}_j^\top - \mathbf{Y}_j\|_F^2, \mathbf{V}_j \rangle.$$

Proposition 8.2.3 thus guarantees that (8.19) has no critical points outside of the consensus subspace. Since we have argued that DGD+LOCAL converges to a second-order critical point $\mathbf{z}^\star$ of (8.19), it follows that $\mathbf{z}^\star$ must be on the consensus subspace; that is, $\mathbf{z}^\star = (\mathbf{U}^{1\star}, \ldots, \mathbf{U}^{J\star}, \mathbf{V}_1^\star, \ldots, \mathbf{V}_J^\star)$ with $\mathbf{U}^{1\star} = \cdots = \mathbf{U}^{J\star} = \mathbf{U}^\star$.

Next, Theorem 8.2.7 guarantees that $\mathbf{z}^\star$ (in which $\mathbf{U}^{1\star} = \cdots = \mathbf{U}^{J\star} = \mathbf{U}^\star$) corresponds to a critical point $(\mathbf{U}^\star, \mathbf{V}^\star)$ of the centralized problem (8.16), which is exactly equivalent to problem (8.14). Here, $\mathbf{V}^\star$ is the concatenation of $\mathbf{V}_1^\star, \ldots, \mathbf{V}_J^\star$ as in (8.15). Theorem 8.3.1 tells us that problem (8.14) has two types of critical points: global minimizers and strict saddles. If $(\mathbf{U}^\star, \mathbf{V}^\star)$ were a strict saddle point of (8.14), Theorem 8.2.7 tells us that $\mathbf{z}^\star$ must then be a strict saddle of (8.19). However, $\mathbf{z}^\star$ is almost surely a second-order critical point of (8.19), where the Hessian has no negative eigenvalues. It follows that $(\mathbf{U}^\star, \mathbf{V}^\star)$ must almost surely be a global minimizer of problem (8.14). $\square$

CHAPTER 9

ALTERNATING MINIMIZATIONS CONVERGE TO SECOND-ORDER OPTIMAL SOLUTIONS

This work[26] studies the second-order convergence for both standard alternating minimization and proximal alternating minimization. We show that under mild assumptions on the (nonconvex) objective function, both algorithms avoid strict saddles almost surely from random initialization. Together with known first-order convergence results, this implies both algorithms converge to a second-order stationary point. This solves an open problem for the second-order convergence of alternating minimization algorithms that have been widely used in practice to solve large-scale nonconvex problems due to their simple implementation, fast convergence, and superb empirical performance.

9.1 Introduction

We consider the following optimization problem over two sets of variables:

$$\underset{\mathbf{x}\in\mathbb{R}^n,\mathbf{y}\in\mathbb{R}^m}{\text{minimize}}\ f(\mathbf{x},\mathbf{y}), \tag{9.1}$$

where $f:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ is a proper continuous (nonconvex) function and the partition of variables into $\mathbf{x}$ and $\mathbf{y}$ blocks typically reflect natural structures within the problem. One approach to solve (9.1) is by concatenating $\mathbf{x}$ and $\mathbf{y}$ as a single variable $\mathbf{z}=(\mathbf{x},\mathbf{y})$ and then directly applying standard iterative algorithms like gradient descent (or its variants) for $f(\mathbf{z})$. Recent progress in nonconvex optimization has provided solid theoretical guarantees for gradient-type algorithms in solving nonconvex problems. In particular, the seminal work [113] shows that gradient descent with random initialization almost surely avoids strict saddles[27] and converges to a second-order critical point. This together with recent results in landscape analysis guarantees that gradient descent can find a global minimum for many popular nonconvex optimization problems, including low-rank matrix recovery [101], matrix completion [100], phase retrieval [215], and deep neural network [216], all of which enjoy a nice landscape that all second-order critical points are global minima.

An alternative approach to solve (9.1) is based on alternating minimization (cf. Algorithm 1, also known as the nonlinear Gauss-Seidel method or the block coordinate descent method) which sequentially optimizes over one variable in each time while keeps the other variable fixed. Compared with gradient-type algorithms, alternating minimization has several advantages: (i) it is easy to implement as there is no need to tune the optimization parameters such as step sizes, (ii) it converges very fast in practice, and (iii) the subproblem is usually easy to solve, e.g., there may exist a closed-form solution. Therefore, alternating minimization has been widely used in many engineering problems.

[26] This is a joint work with Zhihui Zhu and Gongguo Tang [10].
[27] A critical point is a strict saddle if the Hessian at this point has a negative eigenvalue.

Examples include matrix factorization [217, 218], tensor decomposition [75, 219], and the Expectation Maximization (EM) Algorithm [220].

Algorithm 1 Standard Alternating Minimization

1: **Initialization:** $\mathbf{x}_0$.
2: **For** $k = 1, 2, \ldots,$ recursively generate $(\mathbf{x}_k, \mathbf{y}_k)$ by

$$\begin{aligned} \mathbf{y}_k &= \arg\min_{\mathbf{y}\in\mathbb{R}^m} f(\mathbf{x}_{k-1}, \mathbf{y}) \\ \mathbf{x}_k &= \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}, \mathbf{y}_k) \end{aligned} \tag{9.2}$$

Algorithm 2 Proximal Alternating Minimization

1: **Input:** $\beta > L_f$.
2: **Initialization:** $(\mathbf{x}_0, \mathbf{y}_0)$.
3: **For** $k = 1, 2, \ldots,$ recursively generate $(\mathbf{x}_k, \mathbf{y}_k)$ by

$$\begin{aligned} \mathbf{y}_k &= \arg\min_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}) + \frac{\beta}{2}\|\mathbf{y} - \mathbf{y}_{k-1}\|_2^2 \\ \mathbf{x}_k &= \arg\min_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_k) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|_2^2 \end{aligned} \tag{9.3}$$

However, the empirical performance of alternating minimization is not sufficiently substantiated by solid convergence guarantees. In fact, although the idea of alternatingly updating the variables is quite straightforward, the convergence properties for alternating minimization are far more complicated. In particular, alternating minimization may not converge to first-order critical points of the problem [221]. If the function f is strongly bi-convex and satisfies the Kurdyka-Lojasiewicz (KL) property, then Algorithm 1 converges to a critical point of f [222]. The KL property is satisfied by a wide class of nonconvex (and even nonsmooth) functions, including all semi-algebraic functions and sub-analytic functions [223]. To relax the bi-convexity condition, Attouch et al. [223] utilized a proximal method when updating each variable and proved the corresponding proximal alternating minimization (cf. Algorithm 2) converges to a critical point of f as long as f satisfied the KL property. We summarize these results as follows.

Assumption 9.1.1. *f satisfies the KL property and ∇f is Lipschitz continuous on any bounded subset of domain $\mathbb{R}^n \times \mathbb{R}^m$.*

Theorem 9.1.1 (First-order Convergence, [222, 223])**.** *Under Assumption 9.1.1, let $(\mathbf{x}_0, \mathbf{y}_0)$ be any initialization and $(\mathbf{x}_k, \mathbf{y}_k)$ be the sequence generated by Algorithm 1 (if f is further bi-convex) or by Algorithm 2. If the sequence $(\mathbf{x}_k, \mathbf{y}_k)$ is bounded, then it converges to a critical point of f.*

Convergence to a critical point of the objective function alone is not sufficient to explain the successful practical performance of alternating minimization for a considerable body of machine learning problems mentioned above,

which have critical points that are not local minima. Showing the second-order convergence of the alternating minimization methods remains open. The main contribution of this work is closing this gap between the power of alternating minimization in solving nonconvex problems and its second-order convergence. More precisely, we study the second-order convergence of alternating minimizations by answering the following question:

> **Question**: Does (proximal) alternating minimization with a random initialization converge to second-order stationary points with probability one?

We answer this question affirmatively for real analytic functions and establish the following main results on the second-order convergence of Algorithm 1 and Algorithm 2:

Theorem 9.1.2 (Second-order convergence)**.** *Under Assumption 9.1.1, let* $(\mathbf{x}_0, \mathbf{y}_0)$ *be a random initialization and* $(\mathbf{x}_k, \mathbf{y}_k)$ *be the sequence generated by Algorithm 1 (if* f *is further analytic and bi-convex with full-rank cross Hessian at strict saddles) or by Algorithm 2 (if* f *is further bi-smooth). If the sequence* $(\mathbf{x}_k, \mathbf{y}_k)$ *is bounded*[28]*, then it converges to a second-order stationary point of* f *almost surely.*

If additionally, the objective function of the problem satisfies the strict saddle property (i.e., a critical point is either a strict saddle or a local minimum), then Theorem 9.1.2 implies that alternating minimization algorithms with random initialization converge to local minima with probability one. Moreover, many popular machine learning and signal processing problems [98, 100, 101, 224] have no spurious local minimum and thus alternating minimization algorithms converges to a global minimum, partially explaining the good empirical performance of alternating minimization methods in achieving global optimality for these problems.

9.2 Preliminary

Definition 9.2.1. *Let* f *be a twice continuously differentiable function and* ∇ *be the gradient operator. Then we say*

1. $\mathbf{x}$ *is a stationary point (a.k.a. critical point) of* f*, if* $\nabla f(\mathbf{x}) = \mathbf{0}$*;*
2. $\mathbf{x}$ *is a second-order stationary point of* f*, if it is a critical point and* $\nabla^2 f(\mathbf{x})$ *is positive semi-definite;*
3. $\mathbf{x}$ *is a strict saddle of* f*, if it is a critical point and* $\nabla^2 f(\mathbf{x})$ *has at least one negative eigenvalue.*

Definition 9.2.2 (Unstable Fixed Point)**.** *For a mapping* $g : \Omega \to \Omega$*, the set of unstable fixed points is defined as*

$$\mathcal{A}_g = \{\mathbf{x} : g(\mathbf{x}) = \mathbf{x}, \max_i |\lambda_i(Dg(\mathbf{x}))| > 1\},$$

where D *denotes the Jacobian operator.*

[28] The boundedness assumption is automaitcally satisfied if f is coercive, since then the level set $\mathrm{Lev}_f(\mathbf{x}_0, \mathbf{y}_0) := \{(\mathbf{x}, \mathbf{y}) : f(\mathbf{x}, \mathbf{y}) \leq f(\mathbf{x}_0, \mathbf{y}_0)\}$ is bounded for any initialization $(\mathbf{x}_0, \mathbf{y}_0)$ by the coercivity and the nature of (proximal) alternating minimization algorithm ensures that each iterations $(\mathbf{x}, \mathbf{y})$ lying in the level set $\mathrm{Lev}_f(\mathbf{x}_0, \mathbf{y}_0)$ [222, 223].

Theorem 9.2.1 (Theorem 2, [127]). *Let g be a $\mathcal{C}^1$ mapping from Ω to Ω and $\det(Dg(\mathbf{x})) \neq 0$ for all $\mathbf{x} \in \Omega$. Then the set of initial points that converge to unstable fixed points has measure zero, $\mu(\{\mathbf{x}_0 : \lim_{k\to\infty} g^k(\mathbf{x}_0) \in \mathcal{A}_g\}) = 0$. Here $\mu(\cdot)$ denotes the Lebesgue measure.*

Theorem 9.2.1 is instrumental in establishing second-order convergence guarantees for many first-order algorithms in [127]. However, the condition that $\det(Dg(\mathbf{x})) \neq 0$ for all $\mathbf{x} \in \Omega$ is a strongly global property of the Jacobian matrix that is difficult to satisfy and is challenging to verify theoretically. The rest of this section focuses on relaxing this global assumption in Theorem 9.2.1 to a local one so that it can be applied to a larger class of mappings. More precisely, we will replace the global non-singularity condition on the whole domain by a local non-singularity condition around the critical points. This is achieved by refining the arguments used to show Theorem 2 in [127] and the main technical tools are the Zero-Property Theorem and the Maximum Rank Theorem.

Theorem 9.2.2 (Zero-Property Theorem, Theorem 3, [225]). *Let a mapping $g : \Omega \to \Omega$ is continuous and almost everywhere differentiable. Then g satisfies the zero-property (i.e., preimage of any zero-measure set has measure zero) if and only if $\operatorname{rank}(Dg(\mathbf{x})) = \dim(\Omega)$ for almost all $\mathbf{x} \in \Omega$.*

Theorem 9.2.3 (Maximum Rank Theorem, Proposition B.4, [226]). *Suppose $g : \Omega \to \Omega$ is an analytic mapping. $Dg(\mathbf{x})$ achieves the maximum rank almost everywhere in Ω. Here the maximum rank is defined as $\max_{\mathbf{x}\in\Omega} \operatorname{rank}(Dg(\mathbf{x}))$.*

Note the analytic assumption of Theorem 9.2.3 is stronger than infinite differentiability, but still covers a fairly large class of functions, including all elementary functions, most special functions, as well as their combinations and compositions. The Maximum Rank Theorem states that the Jacobian matrix of any analytic mapping almost always achieves the maximum rank. Then as long as the Jacobian matrix is of full-rank at some specific point, the mapping would satisfy the zero-property, which is indicated by Theorem 9.2.2. Now we present the main technical theorem.

Theorem 9.2.4. *Let g be an analytic mapping from Ω to Ω. Then the set of initial points that converge to nondegenerate unstable fixed points has measure zero.*

The proof is adapted from Theorem 2 in [127] and therefore the most important ingredient is the Stable Manifold Theorem Theorem III.7 [227].

Theorem 9.2.5 (Stable Manifold Theorem, Theorem III.7, [227]). *Let $\mathbf{x}^\star$ be a fixed point for a C^r local diffeomorphism $g : U \to E$, where U is a neighborhood of $\mathbf{x}^\star$ in the Banach space E. Suppose that $E = E_s \oplus E_u$, where E_s is the span of the eigenvectors of $Dg(\mathbf{x}^\star)$ corresponding to eigenvalues of magnitude smaller than or equal to 1, and E_u is the span of the eigenvectors of $Dg(\mathbf{x}^\star)$ corresponding to eigenvalues of magnitude larger than 1. Then there exists a C^r embedded disk W^{cs}_{loc} that is tangent to E_s at $\mathbf{x}^\star$ called the local stable center manifold. Moreover, there exists a neighborhood $B_{\mathbf{x}^\star}$ of $\mathbf{x}^\star$, such that $g(W^{cs}_{loc}) \cap B_{\mathbf{x}^\star} \subset W^{cs}_{loc}$ and $\bigcap_{k=0}^{\infty} g^{-k}(B_{\mathbf{x}^\star}) \subset W^{cs}_{loc}$.*

Proof of Theorem 9.2.4. First, for any unstable fixed point $\mathbf{x}^\star \in \mathcal{A}_g$, if it is also non-degenerate, i.e., the Jacobian matrix $Dg(\mathbf{x}^\star)$ is non-singular, then $Dg(\mathbf{x})$ is nonsingular in some neighborhood U of $\mathbf{x}^\star$. This shows $g : U \to g(U)$ is a local diffeomorphism. Then by Stable Manifold Theorem 9.2.5, for any $\mathbf{x}^\star \in \mathcal{A}_g$, there is an associated open neighborhood $B_{\mathbf{x}^\star}$ and thus the union $\bigcup_{\mathbf{x}^\star \in \mathcal{A}_g} B_{\mathbf{x}^\star}$ forms an open cover for $\mathcal{A}_g$. Clearly $\mathcal{A}_g \subset \mathbb{R}^n$, and since $\mathbb{R}^n$ is known to be second-countable (cf. Theorem 10 in [228]), we can extract a countable subcover $\bigcup_{i=1}^\infty B_{\mathbf{x}_i^\star}$ for $\mathcal{A}_g$. Let $W = \{\mathbf{x}_0 \in \Omega : \lim_k g^k(\mathbf{x}_0) \in \mathcal{A}_g\}$. Because $\bigcup_{i=1}^\infty B_{\mathbf{x}_i^\star}$ forms a countable subcover of $\mathcal{A}_g$, $\mathbf{x}^\star \in B_{\mathbf{x}_i^\star}$ for some i, i.e., $\lim_{t\to\infty} g^t(\mathbf{x}_0) \in B_{\mathbf{x}_i^\star}$. That is to say, $g^t(\mathbf{x}_0) \in B_{\mathbf{x}_i^\star}$ for all $t \geq N$ for some sufficiently large N, or equivalently,

$$g^t(\mathbf{x}_0) \in \bigcap_{k=0}^{\infty} g^{-k}(B_{\mathbf{x}_i^\star}) =: S_i \text{ for all } t \geq N.$$

By Stable Manifold Theorem 9.2.5, we have $S_i \subset W_{loc}^{cs}$ with W_{loc}^{cs} of co-dimension at least one (since $\mathbf{x}^\star \in \mathcal{A}_g$). Therefore, S_i has measure zero. Since $g^N(\mathbf{x}_0) \in S_i$ with an unknown non-negative integer N and $\mathbf{x}_0$ is an arbitrary element in W, we must have

$$W \subset \bigcup_{i=1}^{\infty} \bigcup_{N=0}^{\infty} g^{-N}(S_i).$$

Now we show $g^{-N}(S_i)$ has measure zero for any non-negative numbers N and i. Then the proof follows from that any countable union of zero-measure sets has measure zero. Since g is analytic and $\mathbf{x}^\star$ is nondegenerate, i.e., $\operatorname{rank}(Dg(\mathbf{x}^\star)) = n$, which must be the maximum rank of the Jacobian $Dg(\mathbf{x})$ in Ω. Then Theorem 9.2.3 implies that the Jacobian $Dg(\mathbf{x})$ achieves the maximum rank n for almost all $\mathbf{x} \in \Omega$. Further because g is analytic (and hence continuous and almost everywhere differentiable), we can use the Zero-Property Theorem 9.2.2 to get $g^{-N}(S_i)$ has measure zero for all $N \geq 0$. Finally note that the above argument is independent of choice of i. □

9.3 Second-order Convergence of Algorithm 1

For this case when f is strongly bi-convex, we will apply Theorem 9.2.4 to show that Algorithm 1 will not converge to a strict saddle point. Then combining this with the first-order convergence result Theorem 9.1.1, we can get the second-order convergence of Algorithm 1. We first provide some additional assumptions that are used to prove the avoiding-saddle property of Algorithm 1 in solving problem (9.1).

Assumption 9.3.1. *f is a strongly bi-convex[29] analytic function.*

Assumption 9.3.2. *$\nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star, \mathbf{y}^\star)$ has full row rank for all strict saddles $(\mathbf{x}^\star, \mathbf{y}^\star)$.*

Theorem 9.3.1 (Avoiding Strict Saddles)**.** *Suppose f satisfies Assumptions 9.3.1 and 9.3.2. Then solving* (9.1) *using Algorithm 1 with random initialization will not converge to a strict saddle of f almost surely.*

Therefore, together with the first-order convergence Theorem 9.1.1 and noting that any analytic function satisfies Assumption 9.1.1, we have the second-order convergence property of Algorithm 1.

[29] $\nabla^2_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \succ 0$ and $\nabla^2_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \succ 0$ in the whole domain.

Corollary 9.3.1. *Suppose f satisfies Assumptions 9.3.1 and 9.3.2 and the sequence $(\mathbf{x}_k, \mathbf{y}_k)$ generated by Algorithm 1 is bounded. Then solving* (9.1) *using Algorithm 1 with random initialization will converge to a second-order stationary point of f almost surely.*

9.3.1 The Mapping Function of Algorithm 1

First note that Algorithm 1 is well defined under the strong bi-convexity condition in Assumption 9.3.1, since each subproblem minimizes a strongly convex function and thus has a unique optimal solution.

Proposition 9.3.1. *Under Assumption 9.3.1, the following two mappings are well-defined in the whole domain:*

$$\begin{aligned} \boldsymbol{\phi}(\mathbf{x}) &:= \arg\min_{\mathbf{y}\in\mathbb{R}^m} f(\mathbf{x}, \mathbf{y}), \\ \boldsymbol{\psi}(\mathbf{y}) &:= \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}, \mathbf{y}). \end{aligned} \tag{9.4}$$

Proposition 9.3.1 immediately implies Algorithm 1 is well-defined. That is, each subproblem in the k-th iteration has a unique minimizer:

$$\begin{aligned} \mathbf{y}_{k+1} &= \boldsymbol{\phi}(\mathbf{x}_k), \\ \mathbf{x}_{k+1} &= \boldsymbol{\psi}(\mathbf{y}_{k+1}). \end{aligned}$$

By defining the composition $g = \boldsymbol{\psi} \circ \boldsymbol{\phi}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$, we can view the alternating minimization process (9.2) as iteratively performing the following mapping:

$$\mathbf{x}^k = g(\mathbf{x}^{k-1}) = g^k(\mathbf{x}_0) \quad \text{for } k = 1, 2, \ldots \tag{9.5}$$

By the first-order convergence of Algorithm 1, the iterative process (9.5) is continuing until reaching a fixed point $\mathbf{x}^\star$ of the mapping g

$$\mathbf{x}^\star = g(\mathbf{x}^\star). \tag{9.6}$$

In view of (9.4), this is equivalent to

$$\begin{aligned} \mathbf{y}^\star &= \arg\min_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star), \\ \mathbf{x}^\star &= \arg\min_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star) \end{aligned}$$

with $\mathbf{y}^\star := \boldsymbol{\phi}(\mathbf{x}^\star)$. Then together with the strong bi-convexity and the sufficient differentiability (by analytic property) of f, we immediately have that there is a one-to-one correspondence between the fixed points of g and the first-order critical points of f.

Lemma 9.3.1. *A point $\mathbf{x}^\star$ is a fixed point of g if and only if*

$$\nabla f(\mathbf{x}^\star, \mathbf{y}^\star) = \mathbf{0} \tag{9.7}$$

where we have defined $\mathbf{y}^\star = \boldsymbol{\phi}(\mathbf{x}^\star)$ *and* $\nabla f(\mathbf{x}, \mathbf{y}) = [\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})^\top \ \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})^\top]^\top$. *For simplifying notations, we will also often informally write* $\nabla f(\mathbf{x}, \mathbf{y}) = (\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))$.

9.3.2 Proof of Theorem 9.3.1

To use Theorem 9.2.4, we need to show that 1) the mapping g is analytic; 2) all strict saddles of f correspond to unstable fixed points of g; 3) the Jacobian matrix at any strict saddle is full rank. Without loss of generality, we also assume $n \leq m$. This assumption can always be satisfied since otherwise, we can exchange the coordinates of f. We will see this assumption helps to show the non-degenerate property at unstable fixed points of g.

(1) Showing analytic mapping. Towards that end, we derive the closed-form expression of the Jacobian Dg which will also be useful for the remaining proof. To begin, we present an immediate consequence of Proposition 9.3.1.

Proposition 9.3.2. *There exist two well-defined and unique mappings* $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^m$ *and* $\boldsymbol{\psi} : \mathbb{R}^m \to \mathbb{R}^n$ *such that*

$$\begin{aligned} \nabla_{\mathbf{y}} f(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x})) &= \mathbf{0}, \ \ \forall \mathbf{x} \in \mathbb{R}^n, \\ \nabla_{\mathbf{x}} f(\boldsymbol{\psi}(\mathbf{y}), \mathbf{y}) &= \mathbf{0}, \ \ \forall \mathbf{y} \in \mathbb{R}^m. \end{aligned} \tag{9.8}$$

Then we use the Analytic Implicit Function Theorem 9.3.2.

Theorem 9.3.2 (Analytic Implicit Function Theorem, [229], p.34)**.** *Let the function* $h(\mathbf{x}, \mathbf{y}) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ *be analytic. Assume* $h(\mathbf{a}, \mathbf{b}) = \mathbf{0}_m$ *for some point* $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^n \times \mathbb{R}^m$. *If the partial Jacobian* $D_{\mathbf{y}} h(\mathbf{a}, \mathbf{b})$ *is invertible, then there exists an open set* U *of* $\mathbb{R}^n$ *containing* $\mathbf{a}$ *such that there exists a unique analytic function* $\boldsymbol{\phi} : U \to \mathbb{R}^m$ *such that*

$$\boldsymbol{\phi}(\mathbf{a}) = \boldsymbol{b}$$

and

$$h(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x})) = \mathbf{0}_m \quad \textit{for all } \mathbf{x} \in U.$$

Moreover, the Jacobian of $\boldsymbol{\phi}$ *in* U *is given by*

$$D\boldsymbol{\phi}(\mathbf{x}) = -D_{\mathbf{y}} h(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x}))^{-1} D_{\mathbf{x}} h(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x})).$$

We now prove g is analytic.

Lemma 9.3.2. *The mapping* g *is analytic and its Jacobian* $Dg(\mathbf{x})$ *for all* $\mathbf{x} \in \mathbb{R}^n$ *is given by*

$$\begin{aligned} Dg(\mathbf{x}) = &\nabla^2_{\mathbf{x}} f(g(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}))^{-1} \nabla^2_{\mathbf{xy}} f(g(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x})) \times \\ &\nabla^2_{\mathbf{y}} f(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x}))^{-1} \nabla^2_{\mathbf{yx}} f(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x})) \end{aligned} \tag{9.9}$$

Proof of Lemma 9.3.2. From Corollary 9.3.2, we know there have been two well-defined and unique mappings already that satisfies (9.8):

$$\begin{aligned}\nabla_{\mathbf{y}} f(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x})) = \mathbf{0}, \ \ \forall \mathbf{x} \in \mathbb{R}^n \\ \nabla_{\mathbf{x}} f(\boldsymbol{\psi}(\mathbf{y}), \mathbf{y}) = \mathbf{0}, \ \ \forall \mathbf{y} \in \mathbb{R}^m.\end{aligned}$$

Now denote $h_{\mathbf{y}} = \nabla_{\mathbf{y}} f$ and $h_{\mathbf{x}} = \nabla_{\mathbf{x}} f$, which are both analytic as f is analytic. Then the above equations read that

$$\begin{aligned}h_{\mathbf{y}}(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x})) = \mathbf{0}, \quad \forall \mathbf{x} \in \mathbb{R}^n \\ h_{\mathbf{x}}(\boldsymbol{\psi}(\mathbf{y}), \mathbf{y}) = \mathbf{0}, \quad \forall \mathbf{y} \in \mathbb{R}^m.\end{aligned} \tag{9.10}$$

Further note that both $D_{\mathbf{y}} h_{\mathbf{y}} = \nabla^2_{\mathbf{y}} f$ and $D_{\mathbf{x}} h_{\mathbf{x}} = \nabla^2_{\mathbf{x}} f$ are both nonsingular by assumption of strong bi-convexity. Then we can apply Analytic Implicit Function Theorem 9.3.2 to (9.10) to get that $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are the unique analytic mappings satisfying (9.10). Further, using Analytic Implicit Function Theorem 9.3.2, we can compute their Jacobians as

$$\begin{aligned}D\boldsymbol{\phi}(\mathbf{x}) = -\nabla^2_{\mathbf{y}} f(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x}))^{-1} \nabla^2_{\mathbf{xy}} f(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x})); \\ D\boldsymbol{\psi}(\mathbf{y}) = -\nabla^2_{\mathbf{x}} f(\boldsymbol{\psi}(\mathbf{y}), \mathbf{y})^{-1} \nabla^2_{\mathbf{yx}} f(\boldsymbol{\psi}(\mathbf{y}), \mathbf{y}).\end{aligned}$$

Therefore, $g = \boldsymbol{\psi} \circ \boldsymbol{\phi}$ is analytic, as it is a composition of two analytic mappings $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$. Also, the Jacobian Dg is given by the chain rule as follows

$$\begin{aligned}Dg(\mathbf{x}) =& D\boldsymbol{\psi}(\boldsymbol{\phi}(\mathbf{x})) D\boldsymbol{\phi}(\mathbf{x}) \\ =& \nabla^2_{\mathbf{x}} f(g(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}))^{-1} \nabla^2_{\mathbf{xy}} f(g(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x})) \times \\ & \nabla^2_{\mathbf{y}} f(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x}))^{-1} \nabla^2_{\mathbf{yx}} f(\mathbf{x}, \boldsymbol{\phi}(\mathbf{x})).\end{aligned}$$

□

(2) Showing unstable fixed point. First of all, by (9.8), for any strict saddle $(\mathbf{x}^\star, \mathbf{y}^\star)$ of f, $\mathbf{x}^\star = g(\mathbf{x}^\star)$, i.e., $\mathbf{x}^\star$ is a fixed point of g. It remains to show that the maximal magnitude of the eigenvalues of $Dg(\mathbf{x}^\star)$ is greater than 1. Using the fixed point equation $\mathbf{x}^\star = g(\mathbf{x}^\star)$, we first simplify the Jacobian expression (9.9) as

$$\begin{aligned}Dg(\mathbf{x}^\star) =& \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{-1} \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star, \mathbf{y}^\star) \times \\ & \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{-1} \nabla^2_{\mathbf{yx}} f(\mathbf{x}^\star, \mathbf{y}^\star).\end{aligned} \tag{9.11}$$

Define a new matrix

$$\boldsymbol{\Gamma} := \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{1/2} Dg(\mathbf{x}^\star) \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star)^{-1/2}$$

that is similar to $Dg(\mathbf{x}^\star)$. Hence by matrix similarity, they have the same eigenvalues. Plugging $Dg(\mathbf{x}^\star)$ into $\boldsymbol{\Gamma}$, we have

$$
\begin{aligned}
\mathbf{\Gamma} =&(\nabla^2_{\mathbf{x}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{-\frac{1}{2}} \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star,\mathbf{y}^\star) \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{-\frac{1}{2}}) \\
&\times (\nabla^2_{\mathbf{x}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{-\frac{1}{2}} \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star,\mathbf{y}^\star) \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{-\frac{1}{2}})^\top \\
=&\mathbf{L}\mathbf{L}^\top,
\end{aligned}
$$

where $\mathbf{L} := \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{-\frac{1}{2}} \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star,\mathbf{y}^\star) \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{-\frac{1}{2}}$. Therefore, it suffices to show $\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^\top$ has at least an eigenvalue of magnitude greater than 1, since this can imply $Dg(\mathbf{x}^\star)$ has at least an eigenvalue of magnitude greater than 1. Note that $\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^\top$ has at least an eigenvalue of magnitude greater than 1 if and only if the spectral norm of $\|\mathbf{L}\| > 1$.

Now we prove $\|\mathbf{L}\| > 1$ via contradiction. For the sake of contradiction, suppose $\|\mathbf{L}\| \leq 1$. With some standard matrix operations, we can represent Hessian $\nabla^2 f(\mathbf{x}^\star,\mathbf{y}^\star)$ (which is known to have a negative eigenvalue since $(\mathbf{x}^\star,\mathbf{y}^\star)$ is a strict saddle of f) as

$$
\begin{aligned}
&\nabla^2 f(\mathbf{x}^\star,\mathbf{y}^\star) \\
=&\begin{bmatrix} \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star,\mathbf{y}^\star) & \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star,\mathbf{y}^\star) \\ \nabla^2_{\mathbf{yx}} f(\mathbf{x}^\star,\mathbf{y}^\star) & \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star,\mathbf{y}^\star) \end{bmatrix} \\
=&\begin{bmatrix} \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{1/2} & \\ & \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \mathbf{L} \\ \mathbf{L}^\top & \mathbf{I}_m \end{bmatrix} \\
&\begin{bmatrix} \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{1/2} & \\ & \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{1/2} \end{bmatrix}
\end{aligned}
$$

Then we observe that $\begin{bmatrix} \mathbf{I}_n & \mathbf{L} \\ \mathbf{L}^\top & \mathbf{I}_m \end{bmatrix}$ is semi-positive definite:

$$
\begin{aligned}
\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\top \begin{bmatrix} \mathbf{I}_n & \mathbf{L} \\ \mathbf{L}^\top & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} &= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 + 2\mathbf{x}^\top \mathbf{L}\mathbf{y} \\
&\geq \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\|\mathbf{x}\|_2 \|\mathbf{L}\| \|\mathbf{y}\|_2 \\
&\geq \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \geq 0,
\end{aligned}
$$

which holds for all $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$. Consequently, $\nabla^2 f(\mathbf{x}^\star,\mathbf{y}^\star)$ is semi-positive definite, leading to a contradiction. Therefore, we have proved that for any strict saddle $(\mathbf{x}^\star,\mathbf{y}^\star)$, $\mathbf{x}^\star$ is an unstable fixed point of the mapping g.

(3) Showing non-degenerate property. First recall that the Jacobian matrix $Dg(\mathbf{x}^\star)$ at any strict saddles point $\mathbf{x}^\star$ is given by (9.11). Due to the strict positive-definiteness of $\nabla^2_{\mathbf{x}} f(\mathbf{x}^\star,\mathbf{y}^\star)$ and $\nabla^2_{\mathbf{y}} f(\mathbf{x}^\star,\mathbf{y}^\star)$, we know $Dg(\mathbf{x}^\star)$ is similar to a semi-positive definite matrix:

$$
Dg(\mathbf{x}^\star) = \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{1/2} \mathbf{L}\mathbf{L}^\top \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{-1/2}
$$

with $\mathbf{L} = \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{-1/2} \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star,\mathbf{y}^\star) \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star,\mathbf{y}^\star)^{-1/2}$ living in $\mathbb{R}^{n\times m}$. Thus the non-degenerateness immediately follows from Assumption 9.3.2 and the assumption $n \leq m$.

Combining all, we complete the proof of Theorem 9.3.1.

9.3.3 Stylized Application of Algorithm 1

We use a simple example to illustrate our result.

Example 9.3.1 (Best Rank-1 Matrix PCA)**.** *Consider the problem of computing the best rank-1 approximation of a given matrix* $\mathbf{A} \in \mathbb{R}^{n\times m}$ *with* $\mathrm{rank}(\mathbf{A}) = n$*:*

$$f(\mathbf{x},\mathbf{y}) = \frac{1}{2}\|\mathbf{A} - \mathbf{x}\mathbf{y}^\top\|_F^2 + \frac{\lambda}{2}(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2), \tag{9.12}$$

which is an analytic, strongly bi-convex function (cf. Assumption 9.3.1). Note that there are efficient closed-form solutions when using standard alternating minimization Algorithm 1 to solve (9.12)*: given any initialization* $\mathbf{x}_0 \in \mathbb{R}^n$*, the alternating minimization Algorithm 1 recursively generates the following sequence: for* $k = 0, 1, 2, \ldots$

$$\begin{aligned}\mathbf{y}_{k+1} &:= \boldsymbol{\phi}(\mathbf{x}_k) = \frac{\mathbf{A}^\top\mathbf{x}_k}{\lambda + \|\mathbf{x}_k\|_2^2};\\ \mathbf{x}_{k+1} &:= \boldsymbol{\psi}(\mathbf{y}_{k+1}) = \frac{\mathbf{A}\mathbf{y}_{k+1}}{\lambda + \|\mathbf{y}_{k+1}\|_2^2}.\end{aligned}$$

To apply Corollary 9.3.1, one still needs to verify the full-rankness of $\nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star,\mathbf{y}^\star)$ at any strict saddle $(\mathbf{x}^\star,\mathbf{y}^\star)$ of f, where $\mathbf{y}^\star = \boldsymbol{\phi}(\mathbf{x}^\star)$. Direct computations give that

$$\nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star,\mathbf{y}^\star) = 2\mathbf{x}^\star\boldsymbol{\phi}(\mathbf{x}^\star)^\top - \mathbf{A} = \left(2\frac{\mathbf{x}^\star\mathbf{x}^{\star\top}}{\lambda + \|\mathbf{x}^\star\|_2^2} - \mathbf{I}\right)\mathbf{A}$$

Clearly, when $\mathbf{x}^\star = \mathbf{0}$, we have $\nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star,\mathbf{y}^\star) = -\mathbf{A}$ and the full-rankness assumption automatically holds and for $\mathbf{x}^\star \neq \mathbf{0}$, $\mathrm{rank}(\nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star,\mathbf{y}^\star)) = \mathrm{rank}(\mathbf{A})$ provided $\lambda \neq \|\mathbf{x}^\star\|_2^2$. Therefore:

Corollary 9.3.2. *Solving the best rank-1 approximation* (9.12) *for any nonsingular matrix* $\mathbf{A}$*, using Alternating Minimization Algorithm 1 with random initialization, is guaranteed to converge to a second-order stationary point, provided* $\lambda \neq \|\mathbf{x}^\star\|_2^2$*.*

9.4 Second-order Convergence of Algorithm 2

We begin with the following bi-smoothness assumption.

Assumption 9.4.1. $f \in \mathcal{C}^2$ *is* L_f *bi-smooth in the domain, i.e.,* $\max\{\|\nabla^2_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|, \|\nabla^2_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\|\} \leq L_f$ *in the domain.*[30]

In the case where $f(\mathbf{x},\mathbf{y})$ is L_f bi-smooth, we note that Algorithm 2 requires even minor assumptions for it to avoid the strict saddle points.

[30] Any globally smooth function f with $\|\nabla^2 f(\mathbf{x},\mathbf{y})\| \leq L_f$ satisfies Assumption 9.4.1.

Theorem 9.4.1 (Avoiding Strict Saddles)**.** *Suppose f satisfies Assumption 9.4.1. Choose $\beta > L_f$ in Algorithm 2. Then solving* (9.1) *using Algorithm 2 with random initialization will not converge to a strict saddle of f almost surely.*

Therefore, together with the first-order convergence Theorem 9.1.1, we have the second-order convergence property of Algorithm 2.

Corollary 9.4.1. *Suppose f satisfies Assumptions 9.1.1 and 9.4.1 and the sequence $(\mathbf{x}_k, \mathbf{y}_k)$ generated by Algorithm 2 is bounded. Choose $\beta > L_f$ in Algorithm 2. Then solving* (9.1) *using Algorithm 2 with random initialization will return a second-order stationary point of f for almost sure.*

9.4.1 The Mapping Function of Algorithm 2

First from (9.3), we know under the assumptions of $\beta > L_f$ and the L_f bi-smoothness of f, then each subproblem in any iteration of Algorithm 2 is well-defined, since the objective function of each subproblem is strongly convex.

Proposition 9.4.1. *Under Assumption 9.4.1, choose $\beta > L_f$. Then the following two mappings are analytic and well-defined for any $(\mathbf{x}, \mathbf{y})$:*

$$\begin{aligned} \mathbf{p}_\beta(\mathbf{x}, \mathbf{y}) &:= \arg\min_{\mathbf{y}' \in \mathbb{R}^m} f(\mathbf{x}, \mathbf{y}') + \frac{\beta}{2}\|\mathbf{y}' - \mathbf{y}\|_2^2, \\ \mathbf{q}_\beta(\mathbf{x}, \mathbf{y}) &:= \arg\min_{\mathbf{x}' \in \mathbb{R}^n} f(\mathbf{x}', \mathbf{y}) + \frac{\beta}{2}\|\mathbf{x}' - \mathbf{x}\|_2^2. \end{aligned} \tag{9.13}$$

With (9.13), each iteration of Algorithm 2 is equivalent to

$$\begin{aligned} \mathbf{y}_k &= \mathbf{p}_\beta(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}), \\ \mathbf{x}_k &= \mathbf{q}_\beta(\mathbf{x}_{k-1}, \mathbf{y}_k). \end{aligned} \tag{9.14}$$

We define a mapping $g_\beta : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$ such that

$$g_\beta(\mathbf{x}, \mathbf{y}) = \left(\mathbf{q}_\beta(\mathbf{x}, \mathbf{p}_\beta(\mathbf{x}, \mathbf{y})), \mathbf{p}_\beta(\mathbf{x}, \mathbf{y})\right), \tag{9.15}$$

with which we can rewrite (9.14) as

$$\begin{aligned} (\mathbf{x}_k, \mathbf{y}_k) &= (\mathbf{q}_\beta(\mathbf{x}_{k-1}, \mathbf{p}_\beta(\mathbf{x}_{k-1}, \mathbf{y}_{k-1})), \mathbf{p}_\beta(\mathbf{x}_{k-1}, \mathbf{y}_{k-1})) \\ &= g_\beta(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}). \end{aligned}$$

With the implicit function theorem, the following result establishes the expression of the Jacobian matrix for g_β.

Lemma 9.4.1. *For any $(\mathbf{x}, \mathbf{y})$, denote $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = g_\beta(\mathbf{x}, \mathbf{y})$, and assume $\max\{\|\nabla^2_{\mathbf{x}} f(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})\|, \|\nabla^2_{\mathbf{y}} f(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})\|\} \leq L_f$. Set $\beta > L_f$ in Algorithm 2. Then the mapping function g_β is continuous at a neighborhood of $(\mathbf{x}, \mathbf{y})$ and the Jacobian Dg_β is nonsingular at $(\mathbf{x}, \mathbf{y})$ and is given by*

$$Dg_\beta(\mathbf{x},\mathbf{y}) = \begin{bmatrix} \nabla^2_{\mathbf{x}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_n & \nabla^2_{\mathbf{xy}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) \\ \mathbf{0} & \nabla^2_{\mathbf{y}} f(\mathbf{x},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_m \end{bmatrix}^{-1} \\ \begin{bmatrix} \beta \mathbf{I}_n & \mathbf{0} \\ -\nabla^2_{\mathbf{yx}} f(\mathbf{x},\widetilde{\mathbf{y}}) & \beta \mathbf{I}_m \end{bmatrix} \tag{9.16}$$

Proof. Since $\widetilde{\mathbf{y}} = \mathbf{p}_\beta(\mathbf{x},\mathbf{y}), \widetilde{\mathbf{x}} = \mathbf{q}_\beta(\mathbf{x},\widetilde{\mathbf{y}})$, both $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$ can be viewed as functions of $(\mathbf{x},\mathbf{y})$. Note that $(\mathbf{x},\mathbf{y})$ and $(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})$ satisfy the first-order optimality condition of (9.14):

$$\begin{aligned} \nabla_{\mathbf{y}} f(\mathbf{x},\widetilde{\mathbf{y}}) + \beta(\widetilde{\mathbf{y}} - \mathbf{y}) &= \mathbf{0}, \\ \nabla_{\mathbf{x}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) + \beta(\widetilde{\mathbf{x}} - \mathbf{x}) &= \mathbf{0}. \end{aligned} \tag{9.17}$$

We now compute the expression of the Jacobian matrix:

$$Dg_\beta(\mathbf{x},\mathbf{y}) = \begin{bmatrix} \frac{\partial \widetilde{\mathbf{x}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}} & \frac{\partial \widetilde{\mathbf{x}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} \\ \frac{\partial \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}} & \frac{\partial \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} \end{bmatrix}.$$

To obtain the expressions for these partial derivatives $\frac{\partial \widetilde{\mathbf{x}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}}, \frac{\partial \widetilde{\mathbf{x}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}}, \frac{\partial \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}}, \frac{\partial \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}}$, we apply the implicit function theorem to the first-order optimality condition of (9.17) and obtain

$$\begin{aligned} (\nabla^2_{\mathbf{y}} f(\mathbf{x},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_m) \frac{\partial \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}} &= -\nabla^2_{\mathbf{yx}} f(\mathbf{x},\widetilde{\mathbf{y}}), \\ (\nabla^2_{\mathbf{y}} f(\mathbf{x},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_m) \frac{\partial \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} &= \beta \mathbf{I}_m, \\ (\nabla^2_{\mathbf{x}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_n) \frac{\partial \widetilde{\mathbf{x}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}} + \nabla^2_{\mathbf{xy}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) \frac{\partial \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}} &= \beta \mathbf{I}_n, \\ (\nabla^2_{\mathbf{x}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_n) \frac{\partial \widetilde{\mathbf{x}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} + \nabla^2_{\mathbf{xy}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) \frac{\partial \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} &= \mathbf{0}, \end{aligned}$$

which can be rearranged into matrix multiplications as

$$\begin{aligned} &\begin{bmatrix} \nabla^2_{\mathbf{x}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_n & \nabla^2_{\mathbf{xy}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) \\ \mathbf{0} & \nabla^2_{\mathbf{y}} f(\mathbf{x},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \frac{\partial \widetilde{\mathbf{x}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}} & \frac{\partial \widetilde{\mathbf{x}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} \\ \frac{\partial \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}} & \frac{\partial \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} \end{bmatrix} \\ &= \begin{bmatrix} \beta \mathbf{I}_n & \mathbf{0} \\ -\nabla^2_{\mathbf{yx}} f(\mathbf{x},\widetilde{\mathbf{y}}) & \beta \mathbf{I}_m \end{bmatrix} \iff \mathbf{\Gamma}_1 Dg_\beta(\mathbf{x},\mathbf{y}) = \mathbf{\Gamma}_2 \end{aligned}$$

We now show that the matrix $\mathbf{\Gamma}_1$ is nonsingular. Towards that end, suppose there exists $\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}$ such that $\mathbf{\Gamma}_1 \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$, which is equivalent to

$$\begin{aligned} &\begin{bmatrix} (\nabla^2_{\mathbf{x}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_n)\mathbf{u} \\ (\nabla^2_{\mathbf{y}} f(\mathbf{x},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_m)\mathbf{v} \end{bmatrix} = \begin{bmatrix} -\nabla^2_{\mathbf{xy}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})\mathbf{v} \\ \mathbf{0} \end{bmatrix} \\ \iff &\begin{bmatrix} (\nabla^2_{\mathbf{x}} f(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) + \beta \mathbf{I}_n)\mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \iff \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \end{aligned}$$

where we have used the strict positive-definiteness of $\nabla^2_{\mathbf{x}} f(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) + \beta \mathbf{I}_n$ and $\nabla^2_{\mathbf{y}} f(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) + \beta \mathbf{I}_m$ by the assumption. Thus, the matrix $\mathbf{\Gamma}_1$ is nonsingular. Therefore, by the implicit function theorem, $Dg_\beta(\mathbf{x}, \mathbf{y})$ is a continuous function at some neighborhood of $\mathbf{x}, \mathbf{y}$. With similar argument, we obtain that the matrix $\mathbf{\Gamma}_2$ is also nonsingular. Therefore, we have $Dg_\beta(\mathbf{x}, \mathbf{y}) = \mathbf{\Gamma}_1^{-1} \mathbf{\Gamma}_2$ is nonsingular at $\mathbf{x}, \mathbf{y}$. □

9.4.2 Proof of Theorem 9.4.1

We will use Theorem 9.2.1 (a.k.a. Theorem 2 in [127]) to prove Theorem 9.4.1. Therefore, we need to show

1. g_β is a $\mathcal{C}^1$ mapping;
2. $\det(Dg_\beta) \neq 0$ in the whole domain;
3. Any strict saddle of f is an unstable fixed point of g_β.

Showing (1). Because its Jacobian Dg_β is continuous in the whole domain by Lemma 9.4.1 and Assumption 9.4.1.

Showing (2). Because the Jacobian Dg_β is nonsingular in the whole domain by Lemma 9.4.1 and Assumption 9.4.1..

Showing (3). We now show that every strict saddle point $(\mathbf{x}^\star, \mathbf{y}^\star)$ is an unstable fixed point of the mapping. First of all, we show $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a fixed point of g_β. Since a strict saddle point must be a critical point, here we show every critical point of f is a fixed point of g_β. Towards that end, first note that any critical point $(\mathbf{x}, \mathbf{y})$ satisfies $\nabla f(\mathbf{x}, \mathbf{y}) = (\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})) = (\mathbf{0}, \mathbf{0})$, which implies the first optimality condition (9.17). Then noting that Proposition 9.4.1 which states that the mapping g_β is well-defined in the whole domain, we conclude that $(\mathbf{x}, \mathbf{y}) = g_\beta(\mathbf{x}, \mathbf{y})$, i.e., $(\mathbf{x}, \mathbf{y})$ is a fixed point of g_β.

Now we show that the maximum magnitude of eigenvalues of $Dg_\beta(\mathbf{x}^\star, \mathbf{y}^\star)$ is great than 1 at any strict saddle $(\mathbf{x}^\star, \mathbf{y}^\star)$.

Lemma 9.4.2. *Let $(\mathbf{x}^\star, \mathbf{y}^\star)$ be any strict saddle of f with $\max\{\nabla^2_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star), \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star)\} \leq L_f$. Set $\beta > L_f$ in Algorithm 2. Then $\lambda_{\max}\left(Dg_\beta(\mathbf{x}^\star, \mathbf{y}^\star)\right) > 1$, where $\lambda_{\max}$ denotes the largest eigenvalue.*

Proof. To simplify notations, denote

$$\begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} := \begin{bmatrix} \nabla^2_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star) & \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star, \mathbf{y}^\star) \\ \nabla^2_{\mathbf{yx}} f(\mathbf{x}^\star, \mathbf{y}^\star) & \nabla^2_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star) \end{bmatrix}.$$

Then plugging $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = (\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\star, \mathbf{y}^\star)$ to (9.16), we can compute the Jacobian matrix Dg_β at $(\mathbf{x}^\star, \mathbf{y}^\star)$ as

$$\begin{aligned} Dg_\beta(\mathbf{x}^\star, \mathbf{y}^\star) &= \begin{bmatrix} \mathbf{F}_{11} + \beta \mathbf{I}_n & \mathbf{F}_{12} \\ \mathbf{0} & \mathbf{F}_{22} + \beta \mathbf{I}_m \end{bmatrix}^{-1} \begin{bmatrix} \beta \mathbf{I}_n & \mathbf{0} \\ -\mathbf{F}_{21} & \beta \mathbf{I}_m \end{bmatrix} \\ &= \mathbf{I} - \underbrace{\begin{bmatrix} \mathbf{F}_{11} + \beta \mathbf{I}_n & \mathbf{F}_{12} \\ \mathbf{0} & \mathbf{F}_{22} + \beta \mathbf{I}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix}}_{\mathbf{\Phi}} \end{aligned}$$

Therefore, to show that $Dg_\beta(\mathbf{x}^\star, \mathbf{y}^\star)$ has an eigenvalue larger than 1, it suffices to show $\mathbf{\Phi}$ has a negative eigenvalue. We prove this by showing the event that $\det(\mathbf{\Phi} + \mu\mathbf{I}) = 0$ for some $\mu > 0$, where $\det(\cdot)$ denotes the determinant of a matrix. Then with some algebra on the properties of determinant, we have $\det(\mathbf{\Phi} + \mu\mathbf{I}) = 0$ is equivalent to

$$\begin{aligned} &\det\left(\begin{bmatrix}(1+\mu)\mathbf{F}_{11} + \mu\beta\mathbf{I} & (1+\mu)\mathbf{F}_{12} \\ \mathbf{F}_{21} & (1+\mu)\mathbf{F}_{22} + \mu\beta\mathbf{I}\end{bmatrix}\right) = 0 \\ \iff &\det\left(\begin{bmatrix}(1+\mu)\mathbf{F}_{11} + \mu\beta\mathbf{I} & \sqrt{1+\mu}\mathbf{F}_{12} \\ \sqrt{1+\mu}\mathbf{F}_{21} & (1+\mu)\mathbf{F}_{22} + \mu\beta\mathbf{I}\end{bmatrix}\right) = 0 \end{aligned}$$

where the second line has used the property that $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$ and the matrix similarity transform.

Thus, the whole proof now reduces to show that

$$\mathbf{J}(\mu) := \begin{bmatrix}(1+\mu)\mathbf{F}_{11} + \mu\beta\mathbf{I} & \sqrt{1+\mu}\mathbf{F}_{12} \\ \sqrt{1+\mu}\mathbf{F}_{21} & (1+\mu)\mathbf{F}_{22} + \mu\beta\mathbf{I}\end{bmatrix}$$

has a zero eigenvalue for some $\mu > 0$. Note that $\mathbf{J}(\mu)$ is a symmetric matrix (with real eigenvalues) and is a continuous matrix function of μ. Then by Theorem 5.1 in [230], all the eigenvalues of $\mathbf{J}(\mu)$ (including the minimum eigenvalue $\lambda_{\min}(\mathbf{J}(\mu))$) are continuous functions of μ. We will show the real continuous function $\lambda_{\min}(\mathbf{J}(\mu))$ equals zero for some $\mu > 0$. Towards that end, we observe that

$$\begin{aligned} \mathbf{J}(0) &= \begin{bmatrix}\mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22}\end{bmatrix} = \nabla^2 f(\mathbf{x}^\star, \mathbf{y}^\star), \\ \lim_{\mu\to\infty} \frac{\mathbf{J}(\mu)}{\mu} &= \begin{bmatrix}\mathbf{F}_{11} + \beta\mathbf{I} & \\ & \mathbf{F}_{22} + \beta\mathbf{I}\end{bmatrix} \succ 0. \end{aligned}$$

First, since $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a strict saddle of $f(\mathbf{x}, \mathbf{y})$, by definition of strict saddle, we have $\lambda_{\min}(\mathbf{J}(0)) < 0$. Second, since $\beta > L_f \geq \max\{\|\nabla^2_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star)\|, \|\nabla^2_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star)\|\}$ by the assumption, we have both $\mathbf{F}_{11} + \beta\mathbf{I}$ and $\mathbf{F}_{22} + \beta\mathbf{I}_m$ are positive definite and hence $\lambda_{\min}(\mathbf{J}(N)) > 0$ for some sufficiently large N. Finally, since $\lambda_{\min}(\mathbf{J}(\mu))$ is a continuous real-valued function of $\mathbf{X}$, we claim that there must exist a $\mu > 0$ such that $\lambda_{\min}(\mathbf{J}(\mu)) = 0$. □

9.4.3 Stylized Applications of Algorithm 2

We now apply the proximal alternating minimization Algorithm 2 for a popular large-scale matrix optimization problem by the Burer-Monteiro Factorization (BMF) approach [119, 159]: given $\text{minimize}_{\mathbf{M}}\, q(\mathbf{M})$, BMF factorizes $\mathbf{M}$ as $\mathbf{XY}^\top$, and minimizes $f(\mathbf{X}, \mathbf{Y}) := q(\mathbf{XY}^\top)$. It has been shown in [100, 101, 106, 125] that when the original problem q satisfies certain RIP, then any second-order stationary point of f corresponds to a global minimum of q. Therefore, in this sense, the second-order convergence of the proximal alternating algorithm will imply the global optimality convergence. We will focus on two most important matrix problems: matrix sensing and matrix completion.

Example 9.4.1 (Matrix Sensing)**.** *For simplicity, we consider a regularized matrix sensing problem with the objective function* $q_{\mathcal{A}}(\mathbf{M}) = \|\mathcal{A}(\mathbf{M}) - \mathbf{y}\|_2^2 + \lambda\|\mathbf{M}\|_*$ *where* $\mathbf{y}$ *are the observations and* $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ *is the linear sensing*

operator which can be always assumed to have a bounded spectral norm $\|\mathcal{A}\| \leq \mathsf{L}$. The BMF method then solves

$$\underset{\mathbf{X},\mathbf{Y}}{\text{minimize}}\, \|\mathcal{A}(\mathbf{X}\mathbf{Y}^\top) - \mathbf{y}\|_2^2 + \frac{\lambda}{2}(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2) \tag{9.18}$$

Denote $f(\mathbf{X}, \mathbf{Y})$ as the objective function of (9.18)*. Note that in this case, Assumption 9.4.1 is not satisfied since we can not find a universal constant L_f to bound $\|\nabla^2 f(\mathbf{X}, \mathbf{Y})\|$ for all $\mathbf{X}, mY$. However, we note that*

Lemma 9.4.3. *g_β is a forward-invariant mapping on any level set $\Omega := \mathrm{Lev}_f(\mathbf{U}, \mathbf{V})$ for any $\mathbf{U}, \mathbf{V}$, i.e., $g(\Omega) \subseteq \Omega$.*

Proof. In one way, for any $(\mathbf{X}, \mathbf{Y}) \in \Omega$, we have $f(\mathbf{X}, \mathbf{V}) \leq f(\mathbf{U}, \mathbf{V})$ by definition of Ω. In another way, letting $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) = g_\beta(\mathbf{X}, \mathbf{V})$, we have $f(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) \leq f(\mathbf{X}, \mathbf{V})$ by the sufficient decrease property of Algorithm 2 (cf. [223]). Therefore, $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) \in \Omega$. □

Then following the same analysis of [228, Theorem 3], to apply Theorem 9.2.1, it suffices to show:

Proposition 9.4.2. *Choosing $\beta > L_f(\Omega)$ for some constant $L_f(\Omega)$ depending on $\Omega \doteq \mathrm{Lev}_f(\mathbf{U}, \mathbf{V})$, we have: (i) $\det(Dg_\beta) \neq 0$ on Ω, and (ii) all strict saddles of f in Ω are unstable fixed points of g_β. Then by Theorem 10.4.2, the set of all initialization points in Ω that will let g_β converge to strict saddles is of zero Lebesgue measure. Thus together with the first-order convergence (cf. Theorem 9.1.1), Algorithm 2 from random initialization in Ω almost surely converges to a second-order stationary solution of f.*

Proof. With Theorem 10.4.2 and that g_β is forward-invariant in Ω, to prove Proposition 9.4.2, it suffices to show the terms (i) and (ii). To show these two, we first prove a local Lipschitz-gradient condition for f: $\|\nabla^2 f(\mathbf{X}, \mathbf{Y})\| \leq L_f(\Omega)$ for all $(\mathbf{X}, \mathbf{Y}) \in \Omega$. By definition of Ω, $(\mathbf{X}, \mathbf{Y}) \in \Omega$ gives that

$$f(\mathbf{X}, \mathbf{Y}) \leq f(\mathbf{U}, \mathbf{V}) \overset{①}{\Longrightarrow} \begin{cases} \|\mathcal{A}(\mathbf{X}\mathbf{Y}^\top) - \mathbf{y}\|_2^2 & \leq f(\mathbf{U}, \mathbf{V}), \\ \frac{\lambda}{2}\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2 & \leq f(\mathbf{U}, \mathbf{V}). \end{cases}$$

Now deonote $\mathbf{D} \doteq (\mathbf{D}_X, \mathbf{D}_Y)$, $\Lambda \doteq \lambda\|\mathbf{D}\|_F^2$, and compute

$$\begin{aligned} &[\nabla^2 f(\mathbf{X}, \mathbf{Y})](\mathbf{D}, \mathbf{D}) \\ &= 2\|\mathcal{A}(\mathbf{X}\mathbf{D}_Y^\top + \mathbf{D}_X\mathbf{Y}^\top)\|_2^2 + 4\langle \mathcal{A}(\mathbf{D}_X\mathbf{D}_Y^\top), \mathcal{A}(\mathbf{X}\mathbf{Y}^\top) - \mathbf{y}\rangle + \Lambda \\ &\leq \left(4\mathsf{L}^2(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2) + 4\mathsf{L}\|\mathcal{A}(\mathbf{X}\mathbf{Y}^\top) - \mathbf{y}\|_2 + \lambda\right)\|\mathbf{D}\|_F^2. \end{aligned}$$

Together with the definition of spectral norm, this implies

$$\begin{aligned} \|\nabla^2 f(\mathbf{X}, \mathbf{Y})\| &= \underset{\mathbf{D}}{\text{maximize}} [\nabla^2 f(\mathbf{X}\mathbf{Y})](\mathbf{D}, \mathbf{D})/\|\mathbf{D}\|_F^2 \\ &\leq 4\mathsf{L}^2(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2) + 4\mathsf{L}\|\mathcal{A}(\mathbf{X}\mathbf{Y}^\top) - \mathbf{y}\|_2 + \lambda \\ &\leq 8\mathsf{L}^2 f(\mathbf{U}, \mathbf{V})/\lambda + 4\mathsf{L}\sqrt{f(\mathbf{U}, \mathbf{V})} + \lambda \doteq L_f(\Omega), \end{aligned}$$

where the second inequality follows from ①. Now given the local Lipschitz condition in Ω and the forward-invariant property $g(\Omega) \subseteq \Omega$, (i) and (ii) immediately follow from Lemma 9.4.1 and Lemma 9.4.2, respectively. □

Example 9.4.2 (Matrix Completion)**.** *Consider the matrix completion problem which minimizes* $q_{\mathbf{\Omega}}(\mathbf{M}) = \|\mathbf{M} - \mathbf{M}^\star\|_{\mathbf{\Omega}}^2 + \lambda\|\mathbf{M}\|_*$ *with* $\mathbf{M}^\star$ *as the ground-truth,* $\mathbf{\Omega}$ *as the binary mask matrix, and* $\|\mathbf{M}\|_{\mathbf{\Omega}} := \|\mathbf{\Omega} \odot \mathbf{M}\|_F$*. Then the BMF solves*

$$\underset{\mathbf{X},\mathbf{Y}}{\text{minimize}}\, \|\mathbf{X}\mathbf{Y}^\top - \mathbf{M}^\star\|_{\mathbf{\Omega}}^2 + \frac{\lambda}{2}(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2). \tag{9.19}$$

We remark that the same results of Example 9.4.1 (cf. Proposition 9.4.2) can also be applied for the BMF matrix completion (9.19)*, since sensing problem* $q_{\mathcal{A}}(\mathbf{M})$ *reduces to completion propblem* $q_{\mathbf{\Omega}}(\mathbf{M})$ *when choosing the linear sampling operator is a binary sampling operator* $\mathcal{A}(\mathbf{M}) := \mathbf{\Omega} \odot \mathbf{M}$ *and the observations are* $\mathbf{y} := \mathbf{\Omega} \odot \mathbf{M}^\star$*.*

CHAPTER 10

PROVABLE BREGMAN-DIVERGENCE BASED METHODS FOR NONCONVEX AND NON-LIPSCHITZ PROBLEMS

The (global) Lipschitz smoothness condition is crucial in establishing the convergence theory for most optimization methods. Unfortunately, most machine learning and signal processing problems are not Lipschitz smooth. This motivates us to generalize the concept of Lipschitz smoothness condition to the relative smoothness condition, which is satisfied by any finite-order polynomial objective function. Further, this work develops new Bregman-divergence based algorithms that are guaranteed to converge to a second-order stationary point for any relatively smooth problem. In addition, the proposed optimization methods cover both the proximal alternating minimization and the proximal alternating linearized minimization when we specialize the Bregman divergence to the Euclidian distance. Therefore, this work not only develops guaranteed optimization methods for non-Lipschitz smooth problems but also solves an open problem of showing the second-order convergence guarantees for these alternating minimization methods.

10.1 Introduction

Consider minimizing a twice continuously differentiable function

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\mathbf{x}), \tag{10.1}$$

which can be solved by numerous off-the-shelf algorithms, such as first-order methods like vanilla gradient descent (a.k.a. steepest descent), perturbed/stochastic gradient descent, proximal linearized minimization, and nonlinear conjugate gradient method, [113, 114, 126, 127, 148, 212, 231], or second-order methods like the Newton-CG algorithm or proximal quasi-Newton methods [207, 232, 233]. However, all these optimization algorithms require that the gradient of the objective function $f(\mathbf{x})$ should be smooth. In particular, most of the theoretical guarantees for these algorithms require the objective function $f(\mathbf{x})$ to satisfy the *global Lipschitz gradient condition* (a.k.a. second-order Lipschitz condition), that is, there exists a Lipschitz constant $L_f > 0$ such that

$$L_f\mathbf{I} \pm \nabla^2 f(\mathbf{x}) \succeq 0 \tag{10.2}$$

for all $\mathbf{x} \in \mathbb{R}^n$. An immediate consequence of (10.2) is the decent lemma

$$|f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle| \leq \frac{L_f}{2}\|\mathbf{x} - \mathbf{y}\|_2^2, \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{10.3}$$

This decent lemma is central to analyzing the convergence of many iterative algorithms since it guarantees a sufficient decrease of the function value after each iteration. For example, for standard gradient descent with stepsize η,

$$\mathbf{x}^{\ell+1} = g(\mathbf{x}^{\ell}) := \mathbf{x}^{\ell} - \eta \nabla f(\mathbf{x}^{\ell}), \tag{10.4}$$

plugging (10.4) into (10.3) yields the following sufficient decrease property:

$$f(\mathbf{x}^{\ell}) - f(\mathbf{x}^{\ell+1}) \geq \left(\frac{1}{\eta} - \frac{L_f}{2}\right) \|\mathbf{x}^{\ell} - \mathbf{x}^{\ell+1}\|_2^2. \tag{10.5}$$

This sufficient decrease property is a key condition used in the analysis of most (first-order) iterative algorithms to guarantee first-order convergence, i.e., convergence to a first-order stationary point[31].

However, in many applications, e.g., matrix completion [100], phase retrieval [98], matrix sensing [101], and dictionary learning [99], a second-order stationary point is desirable as it is also a global minimum. Although second-order methods like the trust-region algorithm and cubic regularization [170, 234] are guaranteed to converge to a second-order stationary point, their computational complexity is in general much higher than first-order methods. Fortunately, recent work has shown that first-order methods using random initialization or with periodically injected noise can also efficiently avoid strict saddles and converge to a second-order stationary point. In particular, the recent seminal work [113, 127] proves that gradient descent with random initialization almost surely converges to a second-order stationary solution through the so-called Stable Manifold Theorem [227], which suggests that if we view gradient descent in (10.4) as a dynamic system and the iterative mapping function $g(\cdot)$ has a nonsingular Jacobian matrix in the whole domain, then each strict saddle point is unstable and thus the set of initial points that converge to such points has measure zero. The Jacobian matrix of g for the gradient descent algorithm (10.4) is

$$Dg(\mathbf{x}) = \mathbf{I} - \eta \nabla^2 f(\mathbf{x}). \tag{10.6}$$

If f satisfies the global Lipschitz gradient condition (10.2), then $Dg(\mathbf{x}) \succeq (1 - \eta L_f)\mathbf{I}$, which implies that one can always set a sufficiently small stepsize $\eta < 1/L_f$ so that $Dg(\mathbf{x})$ is positive definite (hence nonsingular) in the whole domain. Thus, the global Lipschitz gradient condition is also crucial to [113, 127].

Unfortunately, the objective functions in many machine learning problems—such as low-rank matrix recovery [100, 101], tensor factorization problem [114, 235], neural networks training [55, 236]—do not admit a global Lipschitz gradient constant L_f. This is because for the objective function f to satisfy the global Lipschitz gradient condition (10.2) with constant L_f, all eigenvalues of its Hessian matrix must be upper bounded by L_f in the whole domain (see (10.2)). For that to happen, the objective function should grow at most quadratically. Yet, for matrix factorization and many

[31] We say $\mathbf{x}$ a (first-order) stationary point if $\nabla f(\mathbf{x}) = \mathbf{0}$. We say $\mathbf{x}$ a second-order stationary point if it is a first-order stationary point and the Hessian at this point is positive semi-definite (PSD).

other important problems in practice, the objective functions are higher-order (typically greater than second-order) polynomials, and their Hessian matrices have at least first-order-polynomial entries and thus unbounded eigenvalues over the whole domain. This motivates us to develop new efficient algorithms with the convergence guarantees not requiring the global Lipschitz gradient condition. Therefore, the proposed algorithms can naturally solve these machine learning problems with convergence guarantees.

10.2 Main Results

10.2.1 Beyond Lipschitz Via Bregman Optimizations

Very recently, [237] addressed this longstanding issue through the Bregman distance paradigm, proving that Bregman gradient descent converges to a stationary point of the objective function that is not required to have a globally Lipschitz gradient. Main ingredients of the Bregman distance paradigm are introduced as follows.

Definition 10.2.1 (Bregman Distance)**.** *Given a twice continuously differentiable convex function $h : \mathbb{R}^n \to \mathbb{R}$, the Bregman distance between any $\mathbf{x}$ and $\mathbf{y}$ is defined as*

$$D_h(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}) - h(\mathbf{y}) - \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle. \tag{10.7}$$

For any convex function h, we have $D_h(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and h is called a Bregman distance kernel.

When the Bregman distance kernel is half the squared ℓ_2 norm $h(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, the corresponding Bregman distance reduces to $D_h(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, which is the classical squared Euclidean distance.

Definition 10.2.2 (Adaptive Lipschitz Gradient Condition)**.** *A twice continuously differentiable function f, conveniently denoted as $f \in \mathcal{C}^2$, satisfies the L_f-*adaptive Lipschitz gradient condition *for some Bregman distance kernel $h \in \mathcal{C}^2$ if*

$$L_f \nabla^2 h(\mathbf{x}) \pm \nabla^2 f(\mathbf{x}) \succeq 0 \quad \textit{for all } \mathbf{x} \in \mathbb{R}^n. \tag{10.8}$$

It is worth noting that when $h(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, the adaptive Lipschitz gradient condition (10.8) reduces to $L_f \mathbf{I} \pm \nabla^2 f(\mathbf{x}) \succeq 0$, which is the classical global Lipschitz gradient condition.

When an objective function f satisfies the L_f adaptive Lipschitz gradient condition, a generalized descent lemma

$$|f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle| \leq L_f D_h(\mathbf{x}, \mathbf{y}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \tag{10.9}$$

follows immediately from (10.8). Just as the sufficient descent lemma (10.3) has played a crucial role in deriving first-order convergence theory, this generalized descent lemma (10.9) can be used to obtain the sufficient decrease property of certain Bregman distance-based algorithms (which we define in Section 10.2.3) without the global Lipschitz gradient condition. For example, we establish the following results.

Lemma 10.2.1 (Sufficient Decrease Under Adaptive Lipschitz Gradient Condition)**.** *Suppose $f \in \mathcal{C}^2$ is globally lower-bounded and satisfies the L_f-adaptive Lipschitz gradient condition for some Bregman distance kernel $h \in \mathcal{C}^2$, which is assumed to be σ-strongly convex and super-coercive*[32]*. Then the updating formula* (10.18) *for Bregman gradient descent (Algorithm 5) and* (10.20) *for Bregman proximal minimization (Algorithm 5) are both well-defined and respectively satisfy:*

$$\text{Algorithm 3}: \quad f(\mathbf{x}^{\ell-1}) - f(\mathbf{x}^{\ell}) \geq \left(\frac{1}{\eta} - L_f\right) \frac{\sigma}{2} \|\mathbf{x}^{\ell} - \mathbf{x}^{\ell-1}\|_2^2 \tag{10.10}$$

$$\text{Algorithm 5}: \quad f(\mathbf{x}^{\ell-1}) - f(\mathbf{x}^{\ell}) \geq \frac{\sigma}{2\eta} \|\mathbf{x}^{\ell} - \mathbf{x}^{\ell-1}\|_2^2. \tag{10.11}$$

The proof of Lemma 10.2.1 is in Section G.2. It is worth noting that Bregman gradient descent (Algorithm 3) reduces to standard gradient descent and Bregman proximal minimization (Algorithm 5) reduces to standard proximal minimization when we choose the Bregman distance $D_h(\mathbf{x}, \mathbf{x}^{\ell-1})$ as the classical squared Euclidean distance $\frac{1}{2}\|\mathbf{x} - \mathbf{x}^{\ell-1}\|_2^2$.

10.2.2 Extension to Bregman Alternating Minimizations

Similar results (e.g., Lemma 10.2.2) can be established for *Bregman alternating minimizations* that solve

$$\underset{\mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m}{\text{minimize}} \; f(\mathbf{x}, \mathbf{y}). \tag{10.12}$$

We achieve this by extending the Bregman distance (in Definition 10.2.1) and the adaptive Lipschitz gradient condition (in Definition 10.2.2) into the following double-block versions.

Definition 10.2.3 (Bi-Bregman Distance)**.** *Given a twice continuously differentiable bi-convex function $h(\mathbf{x}, \mathbf{y})$*[33] *from $\mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}$, define the first and second Bregman distances respectively as*

$$D_h^1(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}) = h(\mathbf{x}_1, \mathbf{y}) - h(\mathbf{x}_2, \mathbf{y}) - \langle \nabla_{\mathbf{x}} h(\mathbf{x}_2, \mathbf{y}), \mathbf{x}_1 - \mathbf{x}_2 \rangle, \tag{10.13}$$

$$D_h^2(\mathbf{y}_1, \mathbf{y}_2; \mathbf{x}) = h(\mathbf{x}, \mathbf{y}_1) - h(\mathbf{x}, \mathbf{y}_2) - \langle \nabla_{\mathbf{y}} h(\mathbf{x}, \mathbf{y}_2), \mathbf{y}_1 - \mathbf{y}_2 \rangle \tag{10.14}$$

for any $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and $\mathbf{y}, \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^m$. By the bi-convexity of h, we have both $D_h^1(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}) \geq 0$ and $D_h^2(\mathbf{y}_1, \mathbf{y}_2; \mathbf{x}) \geq 0$ for any $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{y}, \mathbf{y}_1, \mathbf{y}_2$ and h is called a bi-Bregman distance kernel.

Definition 10.2.4 (Bi-Adaptive Lipschitz Gradient Condition)**.** *$f(\mathbf{x}, \mathbf{y}) \in \mathcal{C}^2$ satisfies the* (L_1, L_2)-bi-adaptive Lipschitz gradient condition *for a bi-Bregman distance kernel $h(\mathbf{x}, \mathbf{y}) \in \mathcal{C}^2$ if*

$$L_1 \nabla_{\mathbf{xx}}^2 h(\mathbf{x}, \mathbf{y}) \pm \nabla_{\mathbf{xx}}^2 f(\mathbf{x}, \mathbf{y}) \succeq 0 \text{ and } L_2 \nabla_{\mathbf{yy}}^2 h(\mathbf{x}, \mathbf{y}) \pm \nabla_{\mathbf{yy}}^2 f(\mathbf{x}, \mathbf{y}) \succeq 0, \; \forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^m \tag{10.15}$$

[32] We say h is super-coercive if and only if $\lim_{\|\mathbf{x}\| \to \infty} h(\mathbf{x})/\|\mathbf{x}\|_2 = \infty$ for all $\mathbf{x}$.

[33] We say $h(\mathbf{x}, \mathbf{y})$ is bi-convex if $h(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ for any fixed $\mathbf{y}$ and convex in $\mathbf{y}$ for any fixed $\mathbf{x}$.

Lemma 10.2.2 (Sufficient Decrease Under Bi-Adaptive Lipschitz Gradient Condition)**.** *Suppose $f(\mathbf{x},\mathbf{y}) \in \mathcal{C}^2$ is globally lower-bounded and satisfies the (L_1, L_2)-bi-adaptive Lipschitz gradient condition for some bi-Bregman distance kernel $h(\mathbf{x},\mathbf{y}) \in \mathcal{C}^2$, which is assumed to be σ-strongly bi-convex[34] and bi-super-coercive[35]. Then the updating formula* (10.19) *for Bregman alternating gradient descent (Algorithm 4) and* (10.21) *for Bregman proximal alternating minimization (Algorithm 6) are both well-defined and respectively satisfy[36]:*

$$\text{Algorithm 4}: \quad f(\mathbf{x}^{\ell-1},\mathbf{y}^{\ell-1}) - f(\mathbf{x}^{\ell},\mathbf{y}^{\ell}) \geq \left(\frac{1}{\eta} - L_f\right)\frac{\sigma}{2}\|(\mathbf{x}^{\ell},\mathbf{y}^{\ell}) - (\mathbf{x}^{\ell-1},\mathbf{y}^{\ell-1})\|_2^2 \tag{10.16}$$

$$\text{Algorithm 6}: \quad f(\mathbf{x}^{\ell-1},\mathbf{y}^{\ell-1}) - f(\mathbf{x}^{\ell},\mathbf{y}^{\ell}) \geq \frac{\sigma}{2\eta}\|(\mathbf{x}^{\ell},\mathbf{y}^{\ell}) - (\mathbf{x}^{\ell-1},\mathbf{y}^{\ell-1})\|_2^2 \tag{10.17}$$

The proof of Lemma 10.2.2 follows by noting that $f(\mathbf{x}^{\ell-1},\mathbf{y}^{\ell-1}) - f(\mathbf{x}^{\ell},\mathbf{y}^{\ell}) = f(\mathbf{x}^{\ell-1},\mathbf{y}^{\ell-1}) - f(\mathbf{x}^{\ell},\mathbf{y}^{\ell-1}) + f(\mathbf{x}^{\ell},\mathbf{y}^{\ell-1}) - f(\mathbf{x}^{\ell},\mathbf{y}^{\ell})$ and then recursively applying Lemma 10.2.1 for either fixed $\mathbf{y} = \mathbf{y}^{\ell-1}$ or fixed $\mathbf{x} = \mathbf{x}^{\ell}$.

Note that when the bi-Bregman distance is set as the classical squared Euclidean distance, Bregman alternating gradient descent Algorithm 4 and Bregman proximal alternating minimization Algorithm 6 reduce to proximal alternating linearized minimization [213] and proximal alternating minimization [222, 223], respectively. As a result, our main theory can be applied to remove the requirement of a globally Lipschitz gradient in deriving first-order convergence results for both proximal alternating linearized minimization and proximal alternating minimization. Following the seminal work [113, 127] and using the Stable Manifold Theorem [227], this work also solves an open problem by establishing the second-order convergence of these alternating minimization algorithms. Further, our second-order convergence theories do not require the global Lipschitz gradient condition.

10.2.3 Algorithms

This work will focus on the following four algorithms and derive their second-order convergence theories. Except for Algorithm 3 (cf. [237]), all Algorithms 4- 6 are newly developed and analyzed.

Algorithm 3 Bregman Gradient Descent

1: **Input:** A Bregman kernel h with $L_f\nabla^2 h(\mathbf{x}) \pm \nabla^2 f(\mathbf{x}) \succeq 0$ in the whole domain; Set $\eta \in (0, \frac{1}{L_f})$.
2: **Initialization:** $\mathbf{x}^0$
3: **Recursion:** Iteratively generate a sequence $\{\mathbf{x}^{\ell}\}_{\ell\in\mathbb{N}}$ via

$$\mathbf{x}^{\ell} = \arg\min_{\mathbf{x}} \langle \nabla f(\mathbf{x}^{\ell-1}), \mathbf{x} - \mathbf{x}^{\ell-1}\rangle + \frac{1}{\eta} D_h(\mathbf{x}, \mathbf{x}^{\ell-1}) \tag{10.18}$$

[34] We say $h(\mathbf{x},\mathbf{y})$ is σ-strongly bi-convex if $h(\mathbf{x},\mathbf{y})$ is σ-strongly convex in $\mathbf{x}$ for any fixed $\mathbf{y}$ and is σ-strongly convex in $\mathbf{y}$ for any fixed $\mathbf{x}$.
[35] We say $h(\mathbf{x},\mathbf{y})$ is bi-super-coercive if $\lim_{\|\mathbf{x}\|\to\infty} h(\mathbf{x},\mathbf{y})/\|\mathbf{x}\|_2 = \infty$ and $\lim_{\|\mathbf{y}\|\to\infty} h(\mathbf{x},\mathbf{y})/\|\mathbf{y}\|_2 = \infty$ for all $\mathbf{x},\mathbf{y}$.
[36] We will often use $(\mathbf{a},\mathbf{b}) := [\mathbf{a}^\top\ \mathbf{b}^\top]^\top$.

Algorithm 4 Bregman Alternating Gradient Descent

1: **Input:** A bi-Bregman kernel $h(\mathbf{x},\mathbf{y})$ with both $L_1\nabla^2_{\mathbf{xx}}h(\mathbf{x},\mathbf{y}) \pm \nabla^2_{\mathbf{xx}}f(\mathbf{x},\mathbf{y}) \succeq 0$ and $L_2\nabla^2_{\mathbf{yy}}h(\mathbf{x},\mathbf{y}) \pm \nabla^2_{\mathbf{yy}}f(\mathbf{x},\mathbf{y}) \succeq 0$ in the entire domain; Set $\eta \in (0, \min(\frac{1}{L_1}, \frac{1}{L_2}))$.
2: **Initialization:** $(\mathbf{x}^0, \mathbf{y}^0)$
3: **Recursion:** Iteratively generate a sequence $\{\mathbf{x}^\ell, \mathbf{y}^\ell\}_{\ell\in\mathbb{N}}$ via

$$\begin{aligned}\mathbf{x}^\ell &= \arg\min_{\mathbf{x}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}), \mathbf{x} - \mathbf{x}^{\ell-1}\rangle + \frac{1}{\eta} D_h^1(\mathbf{x}, \mathbf{x}^{\ell-1}; \mathbf{y}^{\ell-1}),\\ \mathbf{y}^\ell &= \arg\min_{\mathbf{y}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}^{\ell}, \mathbf{y}^{\ell-1}), \mathbf{y} - \mathbf{y}^{\ell-1}\rangle + \frac{1}{\eta} D_h^2(\mathbf{y}, \mathbf{y}^{\ell-1}; \mathbf{x}^{\ell})\end{aligned} \tag{10.19}$$

Algorithm 5 Bregman Proximal Minimization

1: **Input:** A Bregman kernel h with $L_f\nabla^2 h(\mathbf{x}) \pm \nabla^2 f(\mathbf{x}) \succeq 0$ in the whole domain; Set $\eta \in (0, \frac{1}{L_f})$.
2: **Initialization:** $\mathbf{x}^0$
3: **Recursion:** Iteratively generate a sequence $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ via

$$\mathbf{x}^\ell = \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{1}{\eta} D_h(\mathbf{x}, \mathbf{x}^{\ell-1}) \tag{10.20}$$

Algorithm 6 Bregman Proximal Alternating Minimization

1: **Input:** A bi-Bregman kernel $h(\mathbf{x},\mathbf{y})$ with both $L_1\nabla^2_{\mathbf{xx}}h(\mathbf{x},\mathbf{y}) \pm \nabla^2_{\mathbf{xx}}f(\mathbf{x},\mathbf{y}) \succeq 0$ and $L_2\nabla^2_{\mathbf{yy}}h(\mathbf{x},\mathbf{y}) \pm \nabla^2_{\mathbf{yy}}f(\mathbf{x},\mathbf{y}) \succeq 0$ in the entire domain; Set $\eta \in (0, \min(\frac{1}{L_1}, \frac{1}{L_2}))$.
2: **Initialization:** $(\mathbf{x}^0, \mathbf{y}^0)$
3: **Recursion:** Iteratively generate a sequence $\{\mathbf{x}^\ell, \mathbf{y}^\ell\}_{\ell\in\mathbb{N}}$ via

$$\begin{aligned}\mathbf{x}^\ell &= \arg\min_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^{\ell-1}) + \frac{1}{\eta} D_h^1(\mathbf{x}, \mathbf{x}^{\ell-1}; \mathbf{y}^{\ell-1}),\\ \mathbf{y}^\ell &= \arg\min_{\mathbf{y}} f(\mathbf{x}^{\ell}, \mathbf{y}) + \frac{1}{\eta} D_h^2(\mathbf{y}, \mathbf{y}^{\ell-1}; \mathbf{x}^{\ell})\end{aligned} \tag{10.21}$$

10.2.4 Main Contributions

Building on the simple and elegant Bregman distance paradigm in [237], we extend their first-order convergence analysis of Bregman gradient descent (Algorithm 3) to a second-order convergence guarantee. In addition, we develop and prove the second-order convergence for Bregman proximal minimization (Algorithm 5), which is a variant of the standard proximal minimization algorithm [212] with the ℓ_2-distance proximal term replaced by the Bregman-distance proximal term.

Furthermore, we generalize the above paradigms to develop new alternating minimization algorithms, including both Bregman alternating gradient descent (Algorithm 4) and Bregman proximal alternating minimization (Algorithm 6). Remarkably, these algorithms are extensions of the standard proximal alternating linearized method [213] and proximal alternating minimization algorithm [222]. It is worth noting 1) that the global Lipschitz gradient condition is required in deriving the second-order convergence of gradient descent and proximal minimization in the

literature [113, 127] and 2) that the second-order convergence of the proximal alternating minimization algorithm is also an open problem [213, 222]. Therefore, this work also complements those works by establishing second-order convergence for proximal alternating minimization [213, 222] when the Bregman distance reduces to the standard Euclidean distance.

In summary, the contributions of this work are mainly in the following two respects.

- First, we develop both gradient-type and proximal-type algorithms through the Bregman distance paradigm to solve the minimization problems

$$\operatorname*{minimize}_{\mathbf{x}} f(\mathbf{x}) \qquad \text{and} \qquad \operatorname*{minimize}_{\mathbf{x},\mathbf{y}} f(\mathbf{x},\mathbf{y}).$$

  Further, all the proposed algorithms are proved to converge to a second-order stationary point of the objective function f without requiring f to have a globally Lipschitz gradient.

- Second, this is the first work to establish second-order convergence results for alternating-minimization type algorithms, as showing the second-order convergence of alternating-minimization type algorithms for nonconvex objective functions is still an open problem. To see this, note that the proximal alternating linearized minimization [213] is a special case of the Bregman alternating gradient descent (cf. Algorithm 4) and that proximal alternating minimization [222] is a particular case of Bregman proximal alternating minimization (Algorithm 6), when we choose the (bi)-Bregman distance as the classical squared Euclidean distance.

We build our main results upon the following assumptions on f and the Bregman kernel h.

Assumption 10.2.1. *$h \in \mathcal{C}^2$ is (bi-)super-coercive and σ-strongly (bi-)convex.*

Assumption 10.2.2. *$f \in \mathcal{C}^2$ is a lower-bounded KL function.*

Assumption 10.2.3. *f satisfies the (bi-)adaptive Lipschitz gradient condition with h.*

Assumption 10.2.4. *The generated sequence $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ (or $\{(\mathbf{x}^\ell,\mathbf{y}^\ell)\}_{\ell\in\mathbb{N}}$) lives in any bounded set $\mathcal{B}$.*

Now we are ready to present the main result, which is proved in Section G.4.

Theorem 10.2.1 (Main Results)**.** *Under Assumptions 10.2.1– 10.2.4, Algorithms 3 and 5 converge almost surely to a second-order stationary point of $f(\mathbf{x})$ in* (10.1) *from random initialization, and Algorithms 4 and 6 converge almost surely to a second-order stationary point of $f(\mathbf{x},\mathbf{y})$ in* (10.12) *from random initialization.*

Remark 10.2.1. Assumption 10.2.2 on f is universal and mild, since KL functions [212] are ubiquitous and include any polynomial function, any ℓ_p norm ($p > 0$ and rational), the ℓ_0 norm, and indicator functions of any semi-algebraic set (see Section 10.4.1 for a detailed discussion). Additionally, the lower-bounded assumption of the objective function is also common in practice, as otherwise the minimization problem would be ill-posed.

Remark 10.2.2. The bounded-sequence assumption (see Assumption 10.2.4) is quite mild since any coercive objective function f satisfies this assumption. It is known that [238, Prop. 11.11] for any coercive function f, its level set $\mathrm{Lev}_f(a) := \{\mathbf{x} : f(\mathbf{x}) \leq a\}$ is bounded for all $a \in \mathbb{R}$. Now by Lemmas 10.2.1 and 10.2.2, all Algorithms 3– 6 can ensure a sufficient decrease of f, implying that all iterations $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ or $\{(\mathbf{x}^\ell, \mathbf{y}^\ell)\}_{\ell\in\mathbb{N}}$ live in the level set $\mathrm{Lev}_f(f(\mathbf{x}^0))$ (or $\mathrm{Lev}_f(f(\mathbf{x}^0, \mathbf{y}^0))$), which is bounded.

Remark 10.2.3. Finally, if the objective function f further satisfies the strict saddle property [99, 114] (i.e., all second-order stationary points are local minimizers), then Theorem 10.2.1 implies that Algorithms 3– 6 almost surely converge to a local minimum from random initialization. Remarkably, many popular (nonconvex) machine learning and signal processing problems [98–101] have no spurious local minima and thus Algorithms 3– 6 converge to a global minimum, implying that global optimality will be achieved when solving these problems.

10.3 Stylized Applications

10.3.1 Polynomial Objective Functions

Many problems of interest involve objective functions that are multi-variate polynomials of certain degrees.

Lemma 10.3.1. *Suppose $f(\mathbf{x})$ (or $f(\mathbf{x}, \mathbf{y})$) is any coercive and lower-bounded dth-degree (or (d_1, d_2)th-degree)[37] polynomial function with $d, d_1, d_2 \geq 2$. Set the Bregman (or bi-Bregman) distance kernel h to be*

$$h(\mathbf{x}) = \frac{\alpha}{d}\|\mathbf{x}\|_2^d + \frac{\sigma}{2}\|\mathbf{x}\|_2^2 + 1 \quad \text{or} \quad h(\mathbf{x}, \mathbf{y}) = (\frac{\alpha}{d_1}\|\mathbf{x}\|_2^{d_1} + \frac{\sigma}{2}\|\mathbf{x}\|_2^2 + 1)(\frac{\alpha}{d_2}\|\mathbf{y}\|_2^{d_2} + \frac{\sigma}{2}\|\mathbf{y}\|_2^2 + 1) \tag{10.22}$$

for any $\alpha, \sigma > 0$. Then $(f(\mathbf{x}), h(\mathbf{x}))$ (or $(f(\mathbf{x}, \mathbf{y}), h(\mathbf{x}, \mathbf{y}))$) satisfies Assumptions 10.2.1– 10.2.4.

Lemma 10.3.1 is proved in Section G.3.1. Now together with Theorem 10.2.1, we obtain that the proposed Bregman algorithms can be used to minimize any lower-bounded finite-degree polynomial.

Corollary 10.3.1. *Suppose $f(\mathbf{x})$ (or $f(\mathbf{x}, \mathbf{y})$) is any coercive and lower-bounded dth-order (or (d_1, d_2)th-order) polynomial function with $d, d_1, d_2 \geq 2$. Set the Bregman (or bi-Bregman) distance kernel h according to* (10.22)*. Then Algorithms 3 and 5 converge almost surely to a second-order stationary point of $f(\mathbf{x})$ in* (10.1) *or $f(\mathbf{w}) := f(\mathbf{x}, \mathbf{y})$ in* (10.12) *when $\mathbf{w} := (\mathbf{x}, \mathbf{y})$ from random initialization, and Algorithms 4 and 6 converge almost surely to a second-order stationary point of $f(\mathbf{x}, \mathbf{y})$ in* (10.12) *from random initialization.*

Recall that the theory for most traditional first-order (or even second-order) and alternating minimization algorithms cannot accommodate high-degree (larger than 2) polynomial objective functions, which sets demanding restrictions on the applications and consequently excludes most interesting practical applications for objective functions in-

[37] We say $f(\mathbf{x})$ is a dth-degree polynomial if the highest order of $\mathbf{x}$ among all monomials of $f(\mathbf{x})$ is d, and $f(\mathbf{x}, \mathbf{y})$ is a (d_1, d_2)th-degree polynomial if the highest order of $\mathbf{x}$ among all monomials of f is d_1 and the highest order of $\mathbf{y}$ among all monomials of f is d_2, where the order of $\mathbf{x}$ of the monomial $x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n} \phi(\mathbf{y})$ for any polynomial function $\phi(\mathbf{y})$ is defined as $\sum_{j=1}^n k_j$ and we define that for $\mathbf{y}$ in a similar way. Note we can view $f(\mathbf{w}) := f(\mathbf{x}, \mathbf{y})$ as a $(d_1 + d_2)$th-degree polynomial of $\mathbf{w} := (\mathbf{x}, \mathbf{y})$.

volving matrix factorizations, which generally involve fourth-degree polynomial objective functions. Corollary 10.3.1 solves this problem by stating that the proposed Algorithms 3– 6 can be applied to any lower-bounded polynomial objective function with provable second-order convergence.

10.3.2 Objective Functions with Polynomial-order Hessian Spectral Norm

Our convergence theory also extends to non-polynomial objective functions as long as the spectral norms of their Hessians have a polynomial growth rate. This is established in Lemma 10.3.2 (which is proved in Section G.3.2) and Corollary 10.3.2.

Lemma 10.3.2. *Suppose $f \in \mathcal{C}^2$ has bounded (partial, resp.) Hessian spectral norms $\|\nabla^2 f(\mathbf{x})\| \leq C_1 + C_2\|\mathbf{x}\|_2^{d-2}$ ($\|\nabla^2_{\mathbf{xx}} f(\mathbf{x},\mathbf{y})\| \leq (C_1+C_2\|\mathbf{x}\|_2^{d_1-2})(C_3+C_4\|\mathbf{y}\|_2^{d_2})$ and $\|\nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y})\| \leq (C_5+C_6\|\mathbf{x}\|_2^{d_1})(C_7+C_8\|\mathbf{y}\|_2^{d_2-2})$, resp.) with $d, d_1, d_2 \geq 2$ and positive constants C_1 to C_8. Set h according to* (10.22)*. Then $(f(\mathbf{x}), h(\mathbf{x}))$ ($(f(\mathbf{x},\mathbf{y}), h(\mathbf{x},\mathbf{y}))$, resp.) satisfies the adaptive (bi-adaptive, resp.) Lipschitz gradient condition.*

Corollary 10.3.2. *Suppose $f \in \mathcal{C}^2$ is any coercive and lower-bounded KL function with its Hessian (or partial Hessian) spectral norms upper bounded by a polynomial as in Lemma 10.3.2. Set h according to* (10.22)*. Then Algorithms 3 and 5 converge almost surely to a second-order stationary point of $f(\mathbf{x})$ in* (10.1) *from random initialization, and Algorithms 4 and 6 converge almost surely to a second-order stationary point of $f(\mathbf{x},\mathbf{y})$ in* (10.12) *from random initialization.*

10.3.3 Burer-Monteiro Factorization Method for Low-rank Matrix Recovery

A popular approach to large-scale matrix optimization problems is the so-called *Burer-Monteiro factorization method* [119, 159]: Given a rank-constrained matrix optimization problem[38]

$$\underset{\mathbf{X}\in\mathbb{S}^n_+ \text{ or } \mathbf{X}\in\mathbb{R}^{n\times m}}{\text{minimize}} \; q(\mathbf{X}) \;\; \text{subject to} \;\; \text{rank}(\mathbf{X}) \leq r, \tag{10.23}$$

the Burer-Monteiro factorization method first parameterizes $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$ (for symmetric case) or $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ (for nonsymmetric case) and then focuses on the new (nonconvex) problem

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\text{minimize}} \, f(\mathbf{U}) := q(\mathbf{U}\mathbf{U}^\top) \quad \text{or} \quad \underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \, f(\mathbf{U},\mathbf{V}) := q(\mathbf{U}\mathbf{V}^\top) \tag{10.24}$$

When the new objective function f is any lower-bounded polynomial or any lower bounded KL function with the Hessian spectral norms upper bounded by a polynomial (which is true in most matrix recovery problems of interest), then the second-order convergence results of Algorithms 3- 6 directly follow from Corollary 10.3.1 and Corollary 10.3.2. More interestingly, when the original objective function $q(\mathbf{X})$ in (10.23) further satisfies $(2r,\delta)$-RIP[39] with $\delta \leq \frac{1}{20}$,

[38] We denote $\mathbb{S}^n_+$ as the set of all $n \times n$ positive semidefinite symmetric matrices.

[39] We say a function $q(\mathbf{X})$ satisfies the $(2r,\delta)$-RIP for some $\delta \in (0,1)$ and positive integer r if for any matrices $\mathbf{X}, \mathbf{M}$ with $\text{rank}(\mathbf{X}) \leq 2r$ and $\text{rank}(\mathbf{M}) \leq 4r$, we have $(1-\delta)\|\mathbf{M}\|_F^2 \leq [\nabla^2 q(\mathbf{X})](\mathbf{M},\mathbf{M}) \leq (1+\delta)\|\mathbf{M}\|_F^2$.

then despite the non-convexity of the new formulated problems (10.24), all second-order stationary points of $f(\mathbf{U})$ (or $f(\mathbf{U}, \mathbf{V})$) correspond to a global minimum of $q(\mathbf{X})$ (cf. [6, 101]). Therefore, we immediately have the following global optimality theory when using the proposed Algorithms 3- 6 to solve (10.24).

Corollary 10.3.3. *Assume $q(\mathbf{X})$ satisfies $(2r, \frac{1}{20})$-RIP. Suppose either (i) f is any lower-bounded finite-degree polynomial; or (ii) $f \in \mathcal{C}^2$ is any coercive and lower-bounded KL function with its (partial) Hessian spectral norms upper bounded by any finite-degree polynomials. Set the (bi-)Bregman distance kernel h according to* (10.22)*. Then applying Algorithms 3 and 5 to* $\operatorname{minimize}_{\mathbf{U}} f(\mathbf{U})$ *in* (10.24) *or applying Algorithms 4 and 6 to* $\operatorname{minimize}_{\mathbf{U},\mathbf{V}} f(\mathbf{U}, \mathbf{V})$ *in* (10.24)*, we can solve* (10.23) *to global optimality almost surely from random initialization.*

10.4 Convergence Analysis

In this section, we first review the main ingredients of the convergence analysis and then use them to prove second-order convergence for Bregman gradient descent Algorithm 3 and Bregman proximal minimization Algorithm 5. Due to the similarity in the proofs of Algorithm 3 and Algorithm 4 (and the proofs of Algorithm 5 and Algorithm 6), we collect the convergence analysis of other Bregman algorithms in Section G.4.

10.4.1 Main Ingredients of First-order Convergence for KL functions

The Kerdyka-Lojasiewicz (KL) property is a characterization of the geometry of an objective function around its critical points, essentially saying that the function landscape is not relatively flat compared with the gradient norm around each critical point. The KL property plays a crucial role in establishing the first-order convergence (a.k.a. sequence convergence) for a number of descent type algorithms (see, e.g., [212, 213, 223, 237, 239]). A function satisfying the KL property is a KL function. KL functions are common in that any proper lower semi-continuous function is a KL function if it is also analytic or semi-algebraic [213, Theorem 5.1]. Therefore, KL functions include but are limited to any polynomial function, any ℓ_p norm ($p > 0$ and rational), the ℓ_0 norm, and indicator functions of any semi-algebraic set. For more discussions and examples, see [212, 213, 223, 239] and their references.

The general framework in [212, 213, 223, 237, 239] uses the KL property to establish the first-order convergence for general descent type algorithms. For this work, we restrict our attention only to twice continuously differentiable functions, which have continuous gradient everywhere in the domain. There are two key ingredients of this framework given in the following definition.

Definition 10.4.1 (Definition 4.1, [237])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. A sequence $\{\mathbf{x}^\ell\}_{\ell \in \mathbb{N}}$ is called a gradient-like descent sequence for f if the following two conditions hold:*

(C1) *Sufficient decrease property: $f(\mathbf{x}^\ell) - f(\mathbf{x}^{\ell+1}) \geq \rho_1 \left\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\right\|_2^2, \ \forall\, \ell \in \mathbb{N}$ for some $\rho_1 > 0$;*

(C2) *Bounded gradient property: $\left\|\nabla f(\mathbf{x}^{\ell+1})\right\|_2 \leq \rho_2 \left\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\right\|_2, \ \forall\, \ell \in \mathbb{N}$ for some $\rho_2 > 0$.*

We note that when $\{\mathbf{x}^\ell\}_{k\in\mathbb{N}}$ is generated by gradient descent with constant stepsize η and the gradient ∇f is globally L_f-Lipschitz, then **(C1)** immediately follows from (10.5) with $\rho_1 = 1/\eta - L_f/2$. It is trivial to using conditions **(C1)** and **(C2)** to show that $\lim_{\ell\to\infty} \|\nabla f(\mathbf{x}^\ell)\|_2 = 0$. However, this is not enough to guarantee the convergence of $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ itself to a unique critical point, due to the possibility of $\mathbf{x}^\ell$ jumping between critical points. This is where the KL property comes to play a role, ensuring a well-behaved geometry of f around each critical point so that such pathological cases will never happen [212, 223].

Theorem 10.4.1 (Theorem 6.2, [237])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function satisfying the KL property. Let $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ be a bounded gradient-like descent sequence for f. Then the sequence $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ converges to a critical point of f.*

10.4.2 Main Ingredients of Second-order Convergence Using Random Initialization

Definition 10.4.2. *Let f be a twice continuously differentiable function and ∇ be the gradient operator. Then*

1. *$\mathbf{x}$ is a first-order stationary point (a.k.a. critical point) of f if the gradient $\nabla f(\mathbf{x}) = \mathbf{0}$;*
2. *$\mathbf{x}$ is a second-order stationary point of f if it is a critical point and $\nabla^2 f(\mathbf{x})$ is positive semi-definite;*
3. *$\mathbf{x}$ is a strict saddle of f if it is a critical point where the Hessian $\nabla^2 f(\mathbf{x})$ has a negative eigenvalue.*

One of the most popular arguments for showing that certain iterative algorithms can almost surely avoid strict saddle points is provided by the seminal work [113, 127], which interprets these algorithms (e.g., gradient descent and proximal minimization) as dynamic systems for which the strict saddle points are unstable fixed points (see Definition 10.4.3) and uses the well-known stable manifold theorem [227] to argue that these will be avoided with high probability.

Definition 10.4.3 (Unstable Fixed Points)**.** *Let g be a $\mathcal{C}^1$ mapping from $\mathcal{X}$ to $\mathcal{X}$. Then the associated set of unstable fixed points is defined as $\mathcal{A}_g = \{\mathbf{x} : g(\mathbf{x}) = \mathbf{x}, \max_i |\lambda_i(Dg(\mathbf{x}))| > 1\}$.*

Theorem 10.4.2 (Theorem 2, [127])**.** *Let g be $\mathcal{C}^1$ mapping and $\det(Dg(\mathbf{x})) \neq 0$ in the entire domain. Then the set of initial points that converge to unstable fixed points has zero measure, $\mu(\{\mathbf{x}^0 : \lim_{\ell\to\infty} g^\ell(\mathbf{x}^0) \in \mathcal{A}_g\}) = 0$. Here $\mu(\cdot)$ counts the Lebesgue measure for a given set.*

Combining this result with the first-order convergence established in Theorem 10.4.1 ensures the desired second-order convergence.

10.4.3 Convergence Analysis of Bregman Gradient Descent

10.4.3.1 First-order Convergence of Algorithm 3

Theorem 10.4.3. *Under Assumptions 10.2.1– 10.2.4, Algorithm 3 must converge to a critical point of f in* (10.1).

Proof. First, it is clear that Algorithm 3 is well-defined in view of Lemma 10.2.1. Then in view of Theorem 10.4.1 and the assumption that f is KL function, it is sufficient to prove that $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ is a gradient-like descent sequence for f (see Definition 10.4.1), i.e., to show:

(C1) Sufficient decrease property: $f(\mathbf{x}^\ell) - f(\mathbf{x}^{\ell+1}) \geq \rho_1 \left\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\right\|_2^2,\ \forall\, \ell \in \mathbb{N}$ for some $\rho_1 > 0$;

(C2) Bounded gradient property: $\left\|\nabla f(\mathbf{x}^{\ell+1})\right\|_2 \leq \rho_2 \left\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\right\|_2,\ \forall\, \ell \in \mathbb{N}$ for some $\rho_2 > 0$.

Condition (C1) follows from (10.10) in Lemma 10.2.1. Condition (C2) holds because by the optimality condition

$$\nabla f(\mathbf{x}^\ell) + (\nabla h(\mathbf{x}^{\ell+1}) - \nabla h(\mathbf{x}^\ell))/\eta = \mathbf{0}, \tag{10.25}$$

we have

$$\|\nabla f(\mathbf{x}^\ell)\|_2 = \frac{1}{\eta}\|\nabla h(\mathbf{x}^{\ell+1}) - \nabla h(\mathbf{x}^\ell)\|_2 \leq \frac{\rho_h(\mathcal{B})}{\eta}\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\|_2,$$

where the inequality follows from Assumption 10.2.4, $h \in \mathcal{C}^2$, and the fact any function in $\mathcal{C}^2$ admits a locally Lipschitz gradient on any bounded set.[40] Therefore, by continuing this argument, we claim that f has a locally $\rho_f(\mathcal{B})$-Lipschitz gradient on $\mathcal{B}$, and we have $\|\nabla f(\mathbf{x}^{\ell+1})\|_2 \leq \left(\frac{\rho_h(\mathcal{B})}{\eta} + \rho_f(\mathcal{B})\right)\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\|_2$. □

10.4.3.2 Second-order Convergence of Algorithm 3

Theorem 10.4.4. *Under Assumptions 10.2.1– 10.2.4, Algorithm 3 with random initialization almost surely converges to a second-order stationary point of f in* (10.1)*.*

Proof. To show the second-order convergence from the first-order convergence, it suffices to show that Algorithm 3 avoids strict saddles. We define (10.18) as $\mathbf{x}^\ell = g(\mathbf{x}^{\ell-1})$ and compute the Jacobian Dg. By the definition of g, we get $Dg(\mathbf{x}^\ell) = \partial\mathbf{x}^{\ell+1}/\partial\mathbf{x}^\ell$. Then we apply the implicit function theorem to the optimality condition (10.25) and in view of the nonsingularity of $\nabla^2 h$, we obtain that Dg is continuous and given by

$$Dg(\mathbf{x}^\ell) = \left[\nabla^2 h(\mathbf{x}^{\ell+1})\right]^{-1}(\nabla^2 h(\mathbf{x}^\ell) - \eta\nabla^2 f(\mathbf{x}^\ell)).$$

Since the above analysis holds for all $\mathbf{x}^\ell \in \mathbb{R}^n$, this further implies that $Dg(\mathbf{x})$ is continuous and given by

$$Dg(\mathbf{x}) = [\nabla^2 h(g(\mathbf{x}))]^{-1}(\nabla^2 h(\mathbf{x}) - \eta\nabla^2 f(\mathbf{x})). \tag{10.26}$$

To show the avoidance of strict saddles, by Theorem 10.4.2, it suffices to show the following conditions:

Showing g is a $\mathcal{C}^1$ mapping. This follows from the continuity of Dg in (10.26).

Showing $\det(Dg) \neq 0$ in the whole domain. By the positive definiteness of $\nabla^2 h$ and $\nabla^2 h \pm \eta\nabla^2 f$,

[40] To see this, for any $h \in \mathcal{C}^2$ and a bounded set $\mathcal{B}$, the Hessian spectral norm $\|\nabla^2 h\|$ (which is a continuous function) must have a maximum $\rho_h(\mathcal{B})$ on the closure of $\mathcal{B}$. This maximum $\rho_h(\mathcal{B})$ can be used as a local Lipschitz gradient constant for h on $\mathcal{B}$.

$$\det(Dg(\mathbf{x})) = \det([\nabla^2 h(g(\mathbf{x}))]^{-1}) \det(\nabla^2 h(\mathbf{x}) - \eta \nabla^2 f(\mathbf{x})) > 0.$$

Showing any strict saddle of f lies in $\mathcal{A}_g$. For any strict saddle $\mathbf{x}^\star$, we have $\mathbf{x}^{\ell+1} = \mathbf{x}^\ell = \mathbf{x}^\star$ satisfies the optimality condition (10.25), so $\mathbf{x}^\star$ is a fixed point, i.e., $g(\mathbf{x}^\star) = \mathbf{x}^\star$. Plugging $g(\mathbf{x}^\star) = \mathbf{x}^\star$ into (10.26):

$$\begin{aligned} Dg(\mathbf{x}^\star) =& [\nabla^2 h(\mathbf{x}^\star)]^{-1}(\nabla^2 h(\mathbf{x}^\star) - \eta \nabla^2 f(\mathbf{x}^\star)) \\ \sim& [\nabla^2 h(\mathbf{x}^\star)]^{-\frac{1}{2}}(\nabla^2 h(\mathbf{x}^\star) - \eta \nabla^2 f(\mathbf{x}^\star))[\nabla^2 h(\mathbf{x}^\star)]^{-\frac{1}{2}} \\ =& \mathbf{I} - \eta [\nabla^2 h(\mathbf{x}^\star)]^{-\frac{1}{2}} \nabla^2 f(\mathbf{x}^\star)[\nabla^2 h(\mathbf{x}^\star)]^{-\frac{1}{2}} := \mathbf{I} - \eta \mathbf{\Phi} \end{aligned}$$

with "$\sim$" denotes the matrix similarity. Therefore, $Dg(\mathbf{x}^\star)$ has an eigenvalue strictly greater than 1 since $\mathbf{\Phi}$ has a negative eigenvalue. This is because $\mathbf{\Phi}$ is congruent to $\nabla^2 f(\mathbf{x}^\star)$, which has a negative eigenvalue. □

10.4.4 Convergence Analysis of Bregman Proximal Minimization

10.4.4.1 First-order Convergence of Algorithm 5

Theorem 10.4.5. *Under Assumptions 10.2.1– 10.2.4, Algorithm 5 must converge to a critical point of f in* (10.1)*.*

Proof. First of all, Algorithm 5 is well-defined in view of Lemma 10.2.1. Then, by Theorem 10.4.1 and the assumption that f is a KL function, it is sufficient to prove that $\{\mathbf{x}^\ell\}_{\ell \in \mathbb{N}}$ is a gradient-like descent sequence for f (see Definition 10.4.1), i.e., to show:

(C1) Sufficient decrease property: $f(\mathbf{x}^\ell) - f(\mathbf{x}^{\ell+1}) \geq \rho_1 \left\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\right\|_2^2, \ \forall\, \ell \in \mathbb{N}$ for some $\rho_1 > 0$;

(C2) Bounded gradient property: $\left\|\nabla f(\mathbf{x}^{\ell+1})\right\|_2 \leq \rho_2 \left\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\right\|_2, \ \forall\, \ell \in \mathbb{N}$ for some $\rho_2 > 0$.

Condition (C1) follows from (10.11) in Lemma 10.2.1. Condition (C2) holds because by the optimality condition

$$\nabla f(\mathbf{x}^{\ell+1}) + (\nabla h(\mathbf{x}^{\ell+1}) - \nabla h(\mathbf{x}^\ell))/\eta = \mathbf{0}, \tag{10.27}$$

we have $\|\nabla f(\mathbf{x}^{\ell+1})\|_2 = \frac{1}{\eta}\|\nabla h(\mathbf{x}^{\ell+1}) - \nabla h(\mathbf{x}^\ell)\|_2 \leq \frac{\rho_h(\mathcal{B})}{\eta}\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\|_2$, where the inequality follows from Assumption 10.2.4, $h \in \mathcal{C}^2$, and Footnote 40. □

10.4.4.2 Second-order Convergence of Algorithm 5

Theorem 10.4.6. *Under Assumptions 10.2.1– 10.2.4, Algorithm 5 with random initialization almost surely converges to a second-order stationary point of f in* (10.1)*.*

Proof. To show the second-order convergence, we define (10.20) as $\mathbf{x}^\ell = g(\mathbf{x}^{\ell-1})$ and compute the Jacobian matrix Dg. By the definition of g, we have $Dg(\mathbf{x}^\ell) = \partial \mathbf{x}^{\ell+1}/\partial \mathbf{x}^\ell$. Now we apply the implicit function theorem to (10.27) and in view of the nonsingularity of $\nabla^2 h + \eta \nabla^2 f$, we obtain that Dg is continuous and given by

$$Dg(\mathbf{x}^\ell) = \left(\nabla^2 h(\mathbf{x}^{\ell+1}) + \eta\nabla^2 f(\mathbf{x}^{\ell+1})\right)^{-1} \nabla^2 h(\mathbf{x}^\ell).$$

Noting that the above argument holds for any $\mathbf{x}^\ell \in \mathbb{R}^n$, we therefore have that $Dg(\mathbf{x})$ is continuous and given by

$$Dg(\mathbf{x}) = \left(\nabla^2 h(g(\mathbf{x})) + \eta\nabla^2 f(g(\mathbf{x}))\right)^{-1} \nabla^2 h(\mathbf{x}). \tag{10.28}$$

By Theorem 10.4.2, to show the mapping g can almost surely avoid the strict saddles, it suffices to show the following conditions:

Showing g is a $\mathcal{C}^1$ mapping. This immediately follows from the continuity of Dg in (10.28).

Showing $\det(Dg) \neq 0$ in the whole domain. Due to the positive definiteness of $\nabla^2 h$ and $\nabla^2 h \pm \eta\nabla^2 f$,

$$\det(Dg(\mathbf{x})) = \det\left(\left[\nabla^2 h(g(\mathbf{x})) + \eta\nabla^2 f(g(\mathbf{x}))\right]^{-1}\right) \det\left(\nabla^2 h(\mathbf{x})\right) > 0.$$

Showing any strict saddle of f lies in $\mathcal{A}_g$. First for any strict saddle $\mathbf{x}^\star$, we have $\mathbf{x}^{\ell+1} = \mathbf{x}^\ell = \mathbf{x}^\star$ satisfies the optimality condition (10.27), indicating $\mathbf{x}^\star$ is a fixed point, i.e., $g(\mathbf{x}^\star) = \mathbf{x}^\star$. Now plugging $g(\mathbf{x}^\star) = \mathbf{x}^\star$ to (10.28), we have

$$\begin{aligned} Dg(\mathbf{x}^\star) =& \left[\nabla^2 h(\mathbf{x}^\star) + \eta\nabla^2 f(\mathbf{x}^\star)\right]^{-1} \nabla^2 h(\mathbf{x}^\star) \\ \sim& [\nabla^2 h(\mathbf{x}^\star) + \eta\nabla^2 f(\mathbf{x}^\star)]^{-1/2} (\nabla^2 h(\mathbf{x}^\star)) [\nabla^2 h(\mathbf{x}^\star) + \eta\nabla^2 f(\mathbf{x}^\star)]^{-1/2} \\ =& \mathbf{I} - \eta [\nabla^2 h(\mathbf{x}^\star) + \eta\nabla^2 f(\mathbf{x}^\star)]^{-1/2} \nabla^2 f(\mathbf{x}^\star) [\nabla^2 h(\mathbf{x}^\star) + \eta\nabla^2 f(\mathbf{x}^\star)]^{-1/2} := \mathbf{I} - \eta\mathbf{\Phi} \end{aligned}$$

where "$\sim$" denotes matrix-similarity. Clearly, we know $Dg(\mathbf{x}^\star)$ has an eigenvalue strictly greater than 1 since $\nabla^2 f(\mathbf{x}^\star)$ has a negative eigenvalue and is congruent to $\mathbf{\Phi}$.

Combining the above three and Theorem 10.4.2, we show that Algorithm 5 can almost surely avoid strict saddles. Finally, combining this with the first-order convergence, we obtain the second-order convergence of Algorithm 5. □

10.5 Conclusion

This work has developed and analyzed four Bregman-type algorithms: Bregman gradient descent, Bregman alternating gradient descent, Bregman proximal minimization, and Bregman proximal alternating minimization. Remarkably, all four algorithms are guaranteed to converge to a second-order stationary point of f where the objective function f is not required to have a globally Lipschitz gradient. Therefore, our result not only improves upon [113, 127] by circumventing the global Lipschitz gradient condition, but also complements [213, 222] by providing second-order convergence for proximal alternating minimization when the Bregman distance reduces to the standard Euclidean distance. Finally, we provide the closed-form updating formula for Bregman (alternating) gradient descent and illustrate the theories using numerical experiments on low-rank matrix factorization in Section G.1.

CHAPTER 11

GENERAL TENSOR RECOVERY VIA ALTERNATING MINIMIZATION

This work studies the problem of retrieving a low-rank tensor under a general linear observation model, including both tensor sensing and tensor completion models. Inspired by the superiority of the matrix nuclear norm in low-rank matrix recovery, we will focus on using tensor nuclear norm to regularize the inverse problem of tensor recovery. Unlike the traditional ways of using approximating values of the tensor nuclear norm due to the NP-hardness of computing the tensor nuclear norm, we use the Burer-Monteiro optimization form of the tensor nuclear norm, and we show this form is tight for any randomly generated tensors. Furthermore, we provide an alternating minimization algorithm to solve the tensor nuclear norm regularized problem, as well as the rigorous mathematical analysis of its global convergence. Our experiments show potential applications of our algorithm and the advantage of our method in term of accuracy and robustness over heuristic approaches.

11.1 Introduction

Tensors can naturally represent massive multi-dimensional data structures arising in many practical applications, which consist of collaborative filtering [53], 3D image processing [52], radar signal processing [54], nonlinear networks design [55, 56] and psychometrics [57]. Tensor methods are the foundations of a lot of machine learning algorithms, including independent component analysis (ICA) [58,59], latent graphical model learning [60], dictionary learning [61], and Gaussian mixture recovery [62].

Despite the utility of tensors in many applications, its widespread adoption in practice has been slow mainly due to two aspects. The first reason is due to the inherent computational intractability when large-size tensors are involved in the algorithms, which is pretty common cases met by modern data applications. The second is due to the lack of simple concepts and available mathematical tools to exploit the inherent low-rankness of the tensor data. Unlike the low-rank matrix recovery based upon the concept of nuclear norm minimization (powered by the singular value decomposition tool SVD) that has earned plenty of attention in the past ten years [64,65,83], however, in low-rank tensor recovery, we are prevented from applying the same nuclear norm regularized idea (cf. [240]) to work on low-rank tensor recovery, which is

$$\underset{\boldsymbol{\mathcal{T}}\in\mathbb{R}^{n_1\times n_2\times n_3}}{\text{minimize}} \ \|\mathbf{y}-\mathrm{A}(\boldsymbol{\mathcal{T}})\|)_2^2+\lambda\|\boldsymbol{\mathcal{T}}\|_* \tag{11.1}$$

where $\|\cdot\|_*$ denotes the tensor nuclear norm and $\boldsymbol{\mathcal{T}}\in\mathbb{R}^{n_1\times n_1\times n_3}$ is the target tensor that we want to recover from the measurement vector $\mathbf{y}\in\mathbb{R}^m$ obtained by a general sampling operator $\mathrm{A}:\mathbb{R}^{n_1\times n_2\times n_3}\to\mathbb{R}^m$. This is mainly owing to the facts that even computing the tensor nuclear norm is an NP-hard problem [66], and there

is a lack of analysis tools for tensor problems. Therefore, the mainstream tensor completion procedures are based upon multiple forms of matricization and utilization of matrix completion to the flattened tensor [52, 84, 85]. As a consequence, when unfolding a three-dimensional tensor into a two-dimensional matrix, the resulting matrix input to the alternating minimization method is typically massive. Further, without tensor nuclear norm as a regularizer, traditional minimization approaches do not yield optimal bounds on the number of measurements required for tensor completion, which is in sharp contrast to the scenario where optimal sample complexity and optimal minimax bound are achieved when using matrix nuclear norm regularizer in low-rank matrix recovery.

To address the above issues, we then get inspirations from the prior work using Burer-Monteiro factorization [91] idea in dealing with matrix nuclear norm regularized problem [92]. Their idea is in two folds. First, they factor data matrix $\mathbf{X}$ into two smaller rectangular matrices $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. Second, they replace the matrix nuclear norm $\|\mathbf{X}\|_*$ as an nonconvex equivalent form $(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)/2$, so that faster and more scalable algorithms can be developed under the new objective function with $(\mathbf{U}, \mathbf{V})$ as the new variables.

In this work, we focus on applying the same Burer-Monteiro factorization idea to low-rank tensor optimization problems for efficiency and scalability purposes. The underlying idea is that for a low-rank tensor, we can always factorize it as the tensor product of three "tall" factor matrices (like what [92] did), which typically have much fewer variables than the original full tensor. Further, we derive a new Burer-Monteiro factorization form of the tensor nuclear norm (cf. Proposition 11.3.1), so that we can incorporate the tensor nuclear norm regularizer into the new factored objective function.

Main Results.

- A straightforward idea of using tensor nuclear norm to do low-rank tensor completion and recovery has been prevented from due to the NP-hardness of computing the tensor nuclear norm [66] and the lack of analysis tools for tensor problems. One main contribution of this work is providing an auxiliary function for the tensor nuclear norm, which is a tight optimization form of the tensor nuclear norm, since we prove that the proposed auxiliary function has the same global optimum as the tensor nuclear norm under the over-parameterization settings(cf. Proposition 11.3.1).

- Another result of this work is developing an alternating-minimization algorithm for solving a general tensor recovery problem, covering both tensor completion and tensor regression. Remarkably, we provide closed-form solutions for each subproblem in the alternating minimization so that an efficient implementation of the algorithm is available. Further, based on the Kurdyka-Łojasiewicz (KL) property (cf. [212, 241]) of the objective function, a rigorous mathematical analysis is provided to guarantee the proposed alternating minimization to globally converge to a stationary point of the axillary tensor nuclear norm regularized problem, with at least a

sub-linear rate of convergence.

- Finally, this work uses the Burer-Monteiro factorization idea to perform the tensor recovery problem. This Burer-Monteiro factorization regime has several favorable properties in performing the massive data operation. In one aspect, it largely reduces the problem dimensions mainly when the involved tensor is large. This is a pretty standard case met in practical applications. In another aspect, using Burer-Monteiro factorization helps to explicitly enforce an upper bound of the tensor rank, so that the recovered tensor is always of low rank.

11.2 General Observation Model

In this work, we focus on third-order nonsymmetric tensors that can be factorized into a linear combination of unit-norm, rank-1 tensors of the form $\mathbf{u} \circ \mathbf{v} \circ \mathbf{w}$, with the (i, j, k)th entry being $u_i v_j w_k$. Assume we have a 3-dimensional data $\boldsymbol{\mathcal{T}}^\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, which admits a low-rank tensor structure, i.e.,

$$\boldsymbol{\mathcal{T}}^\star = \sum_{p=1}^{r} \lambda_p^\star \mathbf{u}_p^\star \circ \mathbf{v}_p^\star \circ \mathbf{w}_p^\star \tag{11.2}$$

where $r \ll \min\{n_1, n_2, n_3\}$ and $\circ$ denotes the tensor/outer product. We assume the coefficients $\{\lambda_p^\star\}$ are always positive, since the sign of any negative coefficient $\lambda_p^\star$ can be absorbed into the factors by noting that $-\mathbf{u}_p^\star \circ \mathbf{v}_p^\star \circ \mathbf{w}_p^\star = (-\mathbf{u}_p^\star) \circ (-\mathbf{v}_p^\star) \circ (-\mathbf{w}_p^\star)$. Further we the tensor factors $\{\mathbf{u}_p^\star, \mathbf{v}_p^\star, \mathbf{w}_p^\star\}$ are locating on the spheres $\|\mathbf{u}_p^\star\|_2 = \|\mathbf{v}_p^\star\|_2 = \|\mathbf{w}_p^\star\|_2 = 1$, since otherwise we can always absorb the lengths of the factors into the coefficients by redefining $\lambda_p^\star = \lambda_p^\star \|\mathbf{u}_p^\star\|_2 \|\mathbf{v}_p^\star\|_2 \|\mathbf{w}_p^\star\|_2$.

This work considers a general linear operator $\mathrm{A}(\cdot) : \mathbb{R}^{n_1 \times n_2 \times n_3} \to \mathbb{R}^m$, which is defined by

$$\mathrm{A}(\boldsymbol{\mathcal{T}}) = \Big\{ \langle \boldsymbol{\mathcal{A}}_p, \boldsymbol{\mathcal{T}} \rangle \Big\}_{p=1}^{m} \text{ for any } \boldsymbol{\mathcal{T}} \in \mathbb{R}^{n_1 \times n_2 \times n_3} \tag{11.3}$$

with m predetermined tensors in $\mathbb{R}^{n_1 \times n_2 \times n_3}$: $\{\boldsymbol{\mathcal{A}}_p\}_{p \in [m]}$.

One equivalent matrix representation of (11.3) is given by

$$\mathrm{A}(\boldsymbol{\mathcal{T}}) = \mathbf{A} \operatorname{vec}(\boldsymbol{\mathcal{T}}),$$

where $\mathbf{A} \in \mathbb{R}^{m \times n_1 n_2 n_3}$ with its pth row being $\mathbf{A}(p, :) = \operatorname{vec}(\boldsymbol{\mathcal{A}}_p)^\top$. The adjoint operator $\mathrm{A}^*(\cdot) : \mathbb{R}^m \to \mathbb{R}^{n_1 \times n_2 \times n_3}$ of $\mathrm{A}(\cdot)$ is given by

$$\mathrm{A}^*(\mathbf{a}) = \sum_{p=1}^{m} a_p \boldsymbol{\mathcal{A}}_p \text{ for any } \mathbf{a} = (a_1, \cdots, a_m).$$

Under this definition, we have $\operatorname{vec}(\mathrm{A}^*(\mathbf{a})) = \mathbf{A}^\top \mathbf{a}$.

The linear operator $\boldsymbol{\mathcal{A}}$ in (11.3) therefore covers almost all linear observation models, including tensor sensing and tensor completion models:

Example 11.2.1 (Tensor Sensing/Regression)**.** *When we observe a small number of linear projections of the ground truth onto some random Gaussian tensors, this linear observation model* (11.3) *includes the Gaussian observation model when* $\{\boldsymbol{\mathcal{A}}_p\}_{p\in[m]}$ *are Gaussian random tensors.*

Example 11.2.2 (Tensor Completion)**.** *This linear observation model* (11.3) *also covers the missing data case if we let* $\{\boldsymbol{\mathcal{A}}_p\}_{p\in[m]}$ *be a subset of cardinality* m *from the canonical basis in* $\mathbb{R}^{n_1\times n_2\times n_3}$*, i.e.,* $\{\mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k\}_{ijk}$*.*

11.3 Tensor Nuclear Norm

For any tensor $\boldsymbol{\mathcal{T}}$, its nuclear norm is defined as [68, Eq. (2.7)].

$$\|\boldsymbol{\mathcal{T}}\|_* = \inf_{\|\mathbf{u}_p\|_2=\|\mathbf{v}_p\|_2=\|\mathbf{w}_p\|_2=1}\left\{\sum_p \lambda_p : \boldsymbol{\mathcal{T}}=\sum_p \lambda_p \mathbf{u}_p\circ\mathbf{v}_p\circ\mathbf{w}_p, \lambda_p>0\right\} \tag{11.4}$$

Therefore, the tensor nuclear norm is the minimal ℓ_1 norm of its expansion coefficients among all valid expansions in terms of unit-norm, rank-1 tensors. The way of defining the tensor nuclear norm is precisely the same as that of defining the matrix nuclear norm. It is known in the literature [4] that the tensor nuclear norm has the identifiability to recover the tensor factors given full measurements of the tensor data. Further, solving tensor nuclear norm is a guarantee to recover the ground-truth tensor factors even in over-complete settings where the number of tensor factors is much larger than the dimension of the tensors [4, Theorem 1.1, Corollary 1.1]. Besides, we expect that tensor nuclear norm, as a particular form of the atomic norm, will achieve the information theoretical limit for tensor completion as the ℓ_1 norm does for compressive sensing, matrix nuclear norm for matrix completion [83], and total variation norm for line spectral estimation with missing data [16].

11.3.1 Burer-Monteiro Optimization Form of Tensor Nuclear Norm

A straightforward approach of applying tensor nuclear norm to low-rank tensor completion and recovery has been limited given that 1) computing the tensor nuclear norm is NP-hard [66]; 2) and unlike matrix problems where a bunch of mathematical tools are available such as eigenvalue decomposition, singular value decomposition, etc., while there is a lack of analysis tools for tensor problems. One main contribution of this work, we provide a new optimization form of tensor nuclear norm based on Burer-Monteiro factorization idea.

Proposition 11.3.1. *Suppose the decomposition that achieves the tensor nuclear norm* $\|\boldsymbol{\mathcal{T}}\|_*$ *involves* r *terms and* $\tilde{r}\geq r$*, then* $\|\boldsymbol{\mathcal{T}}\|_*$ *is equal to the optimal value of the following optimization:*

$$\underset{\{(\mathbf{u}_p,\mathbf{v}_p,\mathbf{w}_p)\}_{p=1}^{\tilde{r}}}{\text{minimize}} \frac{1}{6}\sum_{p=1}^{\tilde{r}}\left[(\|\mathbf{u}_p\|_2^2+\|\mathbf{v}_p\|_2^2\|\mathbf{w}_p\|_2^2)+(\|\mathbf{v}_p\|_2^2+\|\mathbf{u}_p\|_2^2\|\mathbf{w}_p\|_2^2)+(\|\mathbf{w}_p\|_2^2+\|\mathbf{u}_p\|_2^2\|\mathbf{v}_p\|_2^2)\right]$$
$$\text{subject to} \quad \boldsymbol{\mathcal{T}}=\sum_{p=1}^{\tilde{r}}\mathbf{u}_p\circ\mathbf{v}_p\circ\mathbf{w}_p \tag{11.5}$$

Proof. Suppose the tensor nuclear norm $\|\boldsymbol{\mathcal{T}}\|_*$ (see (11.4) for the definition) is achieved by the following decomposition

$$\boldsymbol{\mathcal{T}}=\sum_{p=1}^{r}\lambda_p^\star\mathbf{u}_p^\star\circ\mathbf{v}_p^\star\circ\mathbf{w}_p^\star.$$

Clearly, when

$$\lambda=\frac{\sqrt{\sqrt[3]{2}\left(\sqrt{9\lambda_p^{\star 2}+12}+3\lambda_p^\star\right)^{2/3}-2\sqrt[3]{3}}}{\sqrt[3]{6}\sqrt[6]{\sqrt{9\lambda_p^{\star 2}+12}+3\lambda_p^\star}},$$

we can obtain that $\{\lambda\mathbf{u}_p^\star,\lambda\mathbf{v}_p^\star,\lambda\mathbf{w}_p^\star)\}_{p=1}^{\tilde{r}}$ forms a feasible solution to (11.5) with $\tilde{r}=r$. When $\tilde{r}>r$, we can zero-pad the remaining rank-one factors $\{(\mathbf{u}_p,\mathbf{v}_p,\mathbf{w}_p)\}_{p=r+1}^{\tilde{r}}$. The objective function value at this feasible solution is $\sum_{p=1}^{\tilde{r}}\lambda_p^\star=\|\boldsymbol{\mathcal{T}}\|_*$. This shows that the optimal value of (11.5) is less than or equal to the tensor nuclear norm $\|\boldsymbol{\mathcal{T}}\|_*$.

To show the other direction, suppose an optimal solution of (11.5) is $\{(\mathbf{u}_p,\mathbf{v}_p,\mathbf{w}_p)\}_{p=1}^{\tilde{r}}$. Define

$$a_p=\|\mathbf{u}_p\|_2,\quad b_p=\|\mathbf{v}_p\|_2,\quad c_p=\|\mathbf{w}_p\|_2,\quad \lambda_p=a_pb_pc_p.$$

For p such that $\lambda_p\neq 0$, define

$$\hat{\mathbf{u}}_p=\mathbf{u}_p/a_p,\quad \hat{\mathbf{v}}_p=\mathbf{v}_p/b_p,\quad \hat{\mathbf{w}}_p=\mathbf{w}_p/c_p.$$

Then clearly

$$\boldsymbol{\mathcal{T}}=\sum_{p:\lambda_p\neq 0}\mathbf{u}_p\circ\mathbf{v}_p\circ\mathbf{w}_p=\sum_{p=1}^{\tilde{r}}\lambda_p\hat{\mathbf{u}}_p\circ\hat{\mathbf{v}}_p\circ\hat{\mathbf{w}}_p.$$

Furthermore, by definition (11.5) of tensor nuclear norm, we have $\|\boldsymbol{\mathcal{T}}\|_*\leq\sum_{p=1}^{\tilde{r}}\lambda_p$, and therefore

$$
\begin{aligned}
\|\boldsymbol{\mathcal{T}}\|_* &\leq \frac{1}{3}\sum_{p=1}^{\tilde{r}} a_p b_p c_p + \frac{1}{3}\sum_{p} a_p b_p c_p + \frac{1}{3}\sum_{p=1}^{\tilde{r}} a_p b_p c_p \\
&\leq \frac{1}{3}\sum_{p=1}^{\tilde{r}} (a_p^2/2 + b_p^2 c_p^2/2) + \frac{1}{3}\sum_{p} (b_p^2/2 + a_p^2 c_p^2/2) + \frac{1}{3}\sum_{p} (c_p^2/2 + b_p^2 a_p^2/2) \\
&= \frac{1}{6}\sum_{p=1}^{\tilde{r}} \Big[(\|\mathbf{u}_p\|_2^2 + \|\mathbf{v}_p\|_2^2 \|\mathbf{w}_p\|_2^2) + (\|\mathbf{v}_p\|_2^2 + \|\mathbf{u}_p\|_2^2 \|\mathbf{w}_p\|_2^2) + (\|\mathbf{w}_p\|_2^2 + \|\mathbf{u}_p\|_2^2 \|\mathbf{v}_p\|_2^2) \Big] \\
&= \text{optimal value of (11.5)}.
\end{aligned}
$$

Combining these two directions, we prove that the optimal value of (11.5) is equal to the tensor nuclear norm $\|\boldsymbol{\mathcal{T}}\|_*$.

□

11.4 Alternating Minimization

Similar to using matrix nuclear norm in regularizing matrix inverse problems, the nuclear tensor norm can be used to regularize tensor inverse problems. Assume we observe an unknown low-rank groundtruth tensor $\boldsymbol{\mathcal{T}}^\star$ through the linear observation/measurement model $\mathbf{y} = \mathrm{A}(\boldsymbol{\mathcal{T}}^\star)$, we would like to retrieve the groundtruth tensor $\boldsymbol{\mathcal{T}}^\star$ from the observation $\mathbf{y}$. For instance, when A samples the individual entries of $\boldsymbol{\mathcal{T}}^\star$, we are looking at a tensor completion problem. We propose recovering the low-rank groundtruth tensor $\boldsymbol{\mathcal{T}}^\star$ by solving a tensor nuclear norm regularized least squares Equation (11.1) (cf. [240]). However several difficulties exist in implementing the above method has mentioned in the introduction: 1) Computing the tensor nuclear norm is NP-hard in the worst case; 2) Computational burden is unavoidable when the size of the optimization problem $n_1 n_2 n_3$ is super large.

Therefore, to release the computational burden and bypass the NP-hardness of computing the tensor nuclear norm, we apply the Burer-Monteiro parameterization on the tensor variable $\boldsymbol{\mathcal{T}}$ and the nonconvex reformulation of the tensor nuclear norm eq. (11.5). The resulting program is

$$
\begin{aligned}
\underset{\{(\mathbf{u}_p, \mathbf{v}_p, \mathbf{w}_p)\}_{p=1}^{\tilde{r}}}{\text{minimize}} \ \|\mathbf{y} - \mathrm{A}(\sum_{p=1}^{\tilde{r}} \mathbf{u}_p \circ \mathbf{v}_p \circ \mathbf{w}_p)\|_2^2 + \lambda \sum_{p=1}^{\tilde{r}} \Big[&(\|\mathbf{u}_p\|_2^2 + \|\mathbf{v}_p\|_2^2 \|\mathbf{w}_p\|_2^2) + (\|\mathbf{v}_p\|_2^2 + \|\mathbf{u}_p\|_2^2 \|\mathbf{w}_p\|_2^2) \\
&+ (\|\mathbf{w}_p\|_2^2 + \|\mathbf{u}_p\|_2^2 \|\mathbf{v}_p\|_2^2) \Big]
\end{aligned} \tag{11.6}
$$

Matrix Form. To simplify notations, we denote $\mathbf{U} := [\mathbf{u}_1 \ \cdots \ \mathbf{u}_{\tilde{r}}]$, $\mathbf{V} := [\mathbf{v}_1 \ \cdots \ \mathbf{v}_{\tilde{r}}]$, $\mathbf{W} := [\mathbf{w}_1 \ \cdots \ \mathbf{w}_{\tilde{r}}]$ and reload $\circ$ (when applied to matrices of same columns) as

$$
\mathbf{U} \circ \mathbf{V} \circ \mathbf{W} := \sum_{i=1} \mathbf{U}(:, i) \circ \mathbf{V}(:, i) \circ \mathbf{W}(:, i).
$$

Then the nonconvex formulation (11.6) of the tensor nuclear norm can be rewritten as

$$\underset{\mathbf{U}\in\mathbb{R}^{n_1\times\tilde{r}},\mathbf{V}\in\mathbb{R}^{n_2\times\tilde{r}},\mathbf{W}\in\mathbb{R}^{n_3\times\tilde{r}}}{\text{minimize}} \|\mathbf{y}-\mathrm{A}(\mathbf{U}\circ\mathbf{V}\circ\mathbf{W})\|_2^2$$
$$+\lambda\Big[(\|\mathbf{U}\|_F^2+\mathrm{tr}(\mathbf{V}^\top\mathbf{V}\odot\mathbf{W}^\top\mathbf{W}))+(\|\mathbf{V}\|_F^2+\mathrm{tr}(\mathbf{U}^\top\mathbf{U}\odot\mathbf{W}^\top\mathbf{W}))+(\|\mathbf{W}\|_F^2+\mathrm{tr}(\mathbf{U}^\top\mathbf{U}\odot\mathbf{V}^\top\mathbf{V}))\Big] \tag{11.7}$$

For convenience, we shall denote the objective function of eq. (11.7) as $f(\mathbf{U},\mathbf{V},\mathbf{W})$.

Lemma 11.4.1 (Coordinative strong convexity)**.** *The closed form Hessians $\nabla^2_{\mathbf{U}}f(\cdot),\nabla^2_{\mathbf{V}}f(\cdot),\nabla^2_{\mathbf{W}}f(\cdot)$ is given by*

$$\begin{aligned}
\nabla^2_{\mathbf{U}}f(\mathbf{U},\mathbf{V},\mathbf{W}) &= 2\lambda\mathbf{I}_{n_1}\otimes\mathrm{diag}(\mathbf{1}_r+\mathrm{diag}(\mathbf{V}_{k-1}^\top\mathbf{V}_{k-1}+\mathbf{W}_{k-1}^\top\mathbf{W}_{k-1}))\\
&\quad+2[(\mathbf{W}_{k-1}\circledast\mathbf{V}_{k-1})\otimes\mathbf{I}_{n_1}]^\top\mathbf{A}^\top\mathbf{A}[(\mathbf{W}_{k-1}\circledast\mathbf{V}_{k-1})\otimes\mathbf{I}_{n_1}];\\
\nabla^2_{\mathbf{V}}f(\mathbf{U},\mathbf{V},\mathbf{W}) &= 2\lambda\mathbf{I}_{n_2}\otimes\mathrm{diag}(\mathbf{1}_r+\mathrm{diag}(\mathbf{U}_k^\top\mathbf{U}_k+\mathbf{W}_{k-1}^\top\mathbf{W}_{k-1}))\\
&\quad+2[(\mathbf{W}_{k-1}\circledast\mathbf{U}_k)\otimes\mathbf{I}_{n_2}]^\top\mathbf{A}_{[2,1,3]}^\top\mathbf{A}_{[2,1,3]}[(\mathbf{W}_{k-1}\circledast\mathbf{U}_k)\otimes\mathbf{I}_{n_2}];\\
\nabla^2_{\mathbf{W}}f(\mathbf{U},\mathbf{V},\mathbf{W}) &= 2\lambda\mathbf{I}_{n_3}\otimes\mathrm{diag}(\mathbf{1}_r+\mathrm{diag}(\mathbf{U}_k^\top\mathbf{U}_k+\mathbf{V}_k^\top\mathbf{V}_k))\\
&\quad+2[(\mathbf{U}_k\circledast\mathbf{V}_k)\otimes\mathbf{I}_{n_3}]^\top\mathbf{A}_{[3,2,1]}^\top\mathbf{A}_{[3,2,1]}[(\mathbf{U}_k\circledast\mathbf{V}_k)\otimes\mathbf{I}_{n_3}];
\end{aligned}$$

As a consequence, $f(\mathbf{U},\mathbf{V},\mathbf{W})$ is 2λ-strongly convex with respect to each coordinate ($\mathbf{U},\mathbf{V},\mathbf{W}$) while the other two are fixed. $f(\mathbf{U},\mathbf{V},\mathbf{W})$ is 2λ-strongly convex with respect to each coordinate ($\mathbf{U},\mathbf{V},\mathbf{W}$) while the other two are fixed.

The coordinate strong convexity of the function $f(\mathbf{U},\mathbf{V},\mathbf{W})$ motivates us to solve eq. (11.7) using alternating minimization Algorithm 7.

Algorithm 7 AltMin^λ

1: **Initialization:** $k=1$, λ, and $\mathbf{V}_0\in\mathbb{R}^{n_2\times r},\mathbf{W}_0\in\mathbb{R}^{n_3\times r}$.
2: **while** stop criterion not meet **do**
3: $\quad\mathbf{U}_k=\arg\min_{\mathbf{U}\in\mathbb{R}^{n_1\times r}}f(\mathbf{U},\mathbf{V}_{k-1},\mathbf{W}_{k-1})$;
4: $\quad\mathbf{V}_k=\arg\min_{\mathbf{V}\in\mathbb{R}^{n_2\times r}}f(\mathbf{U}_k,\mathbf{V},\mathbf{W}_{k-1})$;
5: $\quad\mathbf{W}_k=\arg\min_{\mathbf{W}\in\mathbb{R}^{n_3\times r}}f(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W})$;
6: $\quad k=k+1$.
7: **end while**
8: **Output:** factorization $(\mathbf{U}^\lambda,\mathbf{V}^\lambda,\mathbf{W}^\lambda)$.

Lemma 11.4.2. *The iterates sequence $\{(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)\}$ in Algorithm 7 satisfy the following vanishing gradient equations by the first-order optimality condition:*

$$\begin{aligned}
&\nabla_U f(\mathbf{U}_k,\mathbf{V}_{k-1},\mathbf{W}_{k-1})=\mathbf{0},\\
&\nabla_V f(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_{k-1})=\mathbf{0},\\
&\nabla_W f(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)=\mathbf{0},\quad\forall k\geq 1.
\end{aligned}\tag{11.8}$$

By explicitly solving eq. (11.8), we obtain the closed form expressions for $(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)$ in Algorithm 7.

Lemma 11.4.3 (Close-form solutions of Algorithm 7)**.** *The closed form solution exists for each sub-optimization problem in Algorithm 7:*

$$\begin{aligned}
\mathrm{vec}(\mathbf{U}_k) =&(\lambda\mathbf{I}_{n_1} \otimes \mathrm{diag}(\mathbf{1}_r + \mathrm{diag}(\mathbf{V}_{k-1}^\top\mathbf{V}_{k-1} + \mathbf{W}_{k-1}^\top\mathbf{W}_{k-1}))\\
&+ [(\mathbf{W}_{k-1} \circledast \mathbf{V}_{k-1}) \otimes \mathbf{I}_{n_1}]^\top\mathbf{A}^\top\mathbf{A}[(\mathbf{W}_{k-1} \circledast \mathbf{V}_{k-1}) \otimes \mathbf{I}_{n_1}])^{-1}[(\mathbf{W}_{k-1} \circledast \mathbf{V}_{k-1}) \otimes \mathbf{I}_{n_1}]^\top\mathbf{A}^\top\mathbf{y};\\
\mathrm{vec}(\mathbf{V}_k) =&(\lambda\mathbf{I}_{n_2} \otimes \mathrm{diag}(\mathbf{1}_r + \mathrm{diag}(\mathbf{W}_{k-1}^\top\mathbf{W}_{k-1} + \mathbf{U}_k^\top\mathbf{U}_k))\\
&+ [(\mathbf{W}_{k-1} \circledast \mathbf{U}_k) \otimes \mathbf{I}_{n_2}]^\top\mathbf{A}_{[2,1,3]}^\top\mathbf{A}_{[2,1,3]}[(\mathbf{W}_{k-1} \circledast \mathbf{U}_k) \otimes \mathbf{I}_{n_2}])^{-1}[(\mathbf{W}_{k-1} \circledast \mathbf{U}_k) \otimes \mathbf{I}_{n_2}]^\top\mathbf{A}_{[2,1,3]}^\top\mathbf{y};\\
\mathrm{vec}(\mathbf{W}_k) =&(\lambda\mathbf{I}_{n_3} \otimes \mathrm{diag}(\mathbf{1}_r + \mathrm{diag}(\mathbf{V}_k^\top\mathbf{V}_k + \mathbf{U}_{k-1}^\top\mathbf{U}_{k-1}))\\
&+ [(\mathbf{U}_{k-1} \circledast \mathbf{V}_k) \otimes \mathbf{I}_{n_4}]^\top\mathbf{A}_{[3,2,1]}^\top\mathbf{A}_{[3,2,1]}[(\mathbf{U}_{k-1} \circledast \mathbf{V}_k) \otimes \mathbf{I}_{n_4}])^{-1}[(\mathbf{U}_{k-1} \circledast \mathbf{V}_k) \otimes \mathbf{I}_{n_3}]^\top\mathbf{A}_{[3,2,1]}^\top\mathbf{y}
\end{aligned}$$

where $\otimes$ denotes the Kronecker product, $\circledast$ denotes the Khatri-Rao product, and we define for any $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ and any ordered permutation (i_1, i_2, i_3) of $\{1, 2, 3\}$, let $\boldsymbol{\mathcal{T}}_{[i_1,i_2,i_3]}$ be such that

$$\boldsymbol{\mathcal{T}}_{[i_1,i_2,i_3]}(i, j, k) = \boldsymbol{\mathcal{T}}((i, j, k)_{i_1}, (i, j, k)_{i_2}, (i, j, k)_{i_3}).$$

In particular, here, $\mathbf{A}_{[i_1,i_2,i_3]}$ denotes the respective matrix representation with its pth row as $\mathrm{vec}\left(\boldsymbol{\mathcal{A}}_{p[i_1,i_2,i_3]}\right)^\top$.

Remark.

- When $\lambda > 0$, each closed form solution is well-defined by noting that the matrix to be inverted is positive definite with minimum eigenvalue at least λ.
- By definition of $\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k$ from Algorithm 7 and their accessible closed forms from Lemma 11.4.3, we can guarantee that Algorithm 7 generates a decreasing sequence of function values $\{f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ such that $f(\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) \geq f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)$ for any $k \geq 1$.

11.4.1 Boundedness of Variables $\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k$

Definition 11.4.1 (Coercive function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is coercive if $f(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\|_2 \to \infty$.*

Definition 11.4.2 (λ-Strong Coercive function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is λ-strongly coercive if we can decompose f as a sum of nonnegative function g and a λ-scaled squared norm $\lambda\|\mathbf{x}\|_2^2$, that is, $f(\mathbf{x}) = g(\mathbf{x}) + \lambda\|\mathbf{x}\|_2^2$.*

Lemma 11.4.4. *Given a λ-strong coercive function f and a sequence $\{\mathbf{x}_k\}_{k\geq 0}$, if their composition gives a decreasing sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$, then the sequence $\{\mathbf{x}_k\}_{k\geq 0}$ is upper bounded by $\|\mathbf{x}_k\|_2^2 \leq \frac{f(x_0)}{\lambda}$ $\forall k$.*

Proof. From the decreasing property of $f(\mathbf{x}_k)$, we have $f(\mathbf{x}_0) \geq f(\mathbf{x}_1) \geq f(\mathbf{x}_k)$ for all k. By the λ-strong coerciveness of f, we obtain that $f(\mathbf{x}_k) \geq \lambda\|\mathbf{x}_k\|_2^2$. And the proof is completed as $\lambda > 0$. □

An immediate result from Lemma 11.4.4 is that the iterates sequence $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ generated by Algorithm 7 is bounded.

Corollary 11.4.1. *The sequences $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ generated by Algorithm 7 is bounded such that for all k:*

$$\|\mathbf{U}_k\|_F^2 + \|\mathbf{V}_k\|_F^2 + \|\mathbf{W}_k\|_F^2 \leq B^2 \tag{11.9}$$

with $B^2 := \frac{f(\mathbf{U}_0,\mathbf{V}_0,\mathbf{W}_0)}{\lambda}$. Furthermore, each coordinate of the sequence $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ is bounded by B: $\|\mathbf{U}_k\|_F \leq B$, $\|\mathbf{V}_k\|_F \leq B$, $\|\mathbf{W}_k\|_F \leq B$, for all k.

11.4.2 Lipschitz Continuity of Gradient ∇f along Solution Path

In this part, we will use the boundedness of $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\geq 0}$ to show that the objective function f in (11.7) is Lipschitz smooth along the solution path $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\geq 0}$. In precise, we will bound the Lipschitz constant of its gradient ∇f along the solution path, which is equivalent to bound the spectral norm of its Hessian $\nabla^2 f$.

The following lemma shows that our objective in (11.7) is C^1 smooth on any bounded subset.

Lemma 11.4.5 (Lipchitz continuity)**.** *The objective function $f(\mathbf{U}, \mathbf{V}, \mathbf{W})$ has Lipschitz continuous gradient with the Lipschitz constant as*

$$\mathcal{L}_g := 15\|\mathbf{A}\|_2^2 B^4 + 6\|\mathbf{A}\|_2\|\mathbf{y}\|_2 B + \lambda(1 + 3B^2)$$

in any bounded ℓ_2-norm ball $\{(\mathbf{U}, \mathbf{V}) : \|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 \leq B\}$ for the ℓ_2 norm ball $\mathcal{B}(\mathbf{0}, B)$ of radius B (11.9).

Proof. We first compute the spectral norm of the Hessian is through its quadratic form: let $\mathbf{D} := \begin{bmatrix}\mathbf{D}_U^\top & \mathbf{D}_V^\top & \mathbf{D}_W^\top\end{bmatrix}^\top$ with $\mathbf{D}_U \in \mathbb{R}^{n_1\times r}$, $\mathbf{D}_V \in \mathbb{R}^{n_2\times r}$, $\mathbf{D}_W \in \mathbb{R}^{n_3\times r}$, then the Hessian directional quadratic form of f along $\mathbf{D}$ is given by

$$\begin{aligned}
&[\nabla^2 f(\mathbf{U}, \mathbf{V}, \mathbf{W})](\mathbf{D}, \mathbf{D}) \\
=&2\langle \mathrm{A}^*(\mathrm{A}(\mathbf{U}\circ\mathbf{V}\circ\mathbf{W}) - \mathbf{y}), \mathbf{D}_U\circ\mathbf{D}_V\circ\mathbf{W} + \mathbf{D}_U\circ\mathbf{V}\circ\mathbf{D}_W + \mathbf{U}\circ\mathbf{D}_V\circ\mathbf{D}_W\rangle && (\Pi_1)\\
&+ \langle \mathrm{A}^*\mathrm{A}(\mathbf{D}_U\circ\mathbf{V}\circ\mathbf{W} + \mathbf{U}\circ\mathbf{D}_V\circ\mathbf{W} + \mathbf{U}\circ\mathbf{V}\circ\mathbf{D}_W), \mathbf{D}_U\circ\mathbf{V}\circ\mathbf{W} + \mathbf{U}\circ\mathbf{D}_V\circ\mathbf{W} + \mathbf{U}\circ\mathbf{V}\circ\mathbf{D}_W\rangle && (\Pi_2)\\
&+ \lambda(\|\mathbf{D}_U\|_F^2 + \|\mathbf{D}_V\|_F^2 + \|\mathbf{D}_W\|_F^2) && (\Pi_3)\\
&+ \lambda\,\mathrm{tr}(\mathbf{D}_U^\top\mathbf{D}_U \odot (\mathbf{V}^\top\mathbf{V} + \mathbf{W}^\top\mathbf{W}) + \mathbf{D}_V^\top\mathbf{D}_V \odot (\mathbf{U}^\top\mathbf{U} + \mathbf{W}^\top\mathbf{W}) + \mathbf{D}_W^\top\mathbf{D}_W \odot (\mathbf{U}^\top\mathbf{U} + \mathbf{V}^\top\mathbf{V})) && (\Pi_4)\\
&+ 4\lambda\,\mathrm{tr}(\mathbf{U}^\top\mathbf{D}_U \odot \mathbf{V}^\top\mathbf{D}_V + \mathbf{U}^\top\mathbf{D}_U \odot \mathbf{W}^\top\mathbf{D}_W + \mathbf{V}^\top\mathbf{D}_V \odot \mathbf{W}^\top\mathbf{D}_W) && (\Pi_5)
\end{aligned}$$

Since the spectral norm of the Hessian $\nabla^2 f(\mathbf{U}, \mathbf{V}, \mathbf{W})$ is given by the maximum of the Hessian quadratic form $[\nabla^2 f(\mathbf{U}, \mathbf{V}, \mathbf{W})](\mathbf{D}, \mathbf{D})$ for all normalized $\mathbf{D}$, we then upper bound this Hessian quadratic form $[\nabla^2 f(\mathbf{U}, \mathbf{V}, \mathbf{W})](\mathbf{D}, \mathbf{D})$ by controlling Π_1 to Π_5 through Lemma 11.4.6.

Lemma 11.4.6. *For any matrices $\mathbf{X}, \mathbf{W}, \mathbf{Y}, \mathbf{Z}$ that have the same number of columns, we can easily verify that the following holds:*

1. $\|\mathbf{X}\circ\mathbf{Y}\circ\mathbf{Z}\|_F \leq \|\mathbf{X}\|_F\|\mathbf{Y}\|_F\|\mathbf{Z}\|_F$
2. $\mathrm{tr}(\mathbf{X}^\top\mathbf{X} \odot \mathbf{Y}^\top\mathbf{Y}) \leq \|\mathbf{X}\|_F^2\|\mathbf{Y}\|_F^2$

3. $\mathrm{tr}(\mathbf{X}^\top\mathbf{Y}\odot\mathbf{W}^\top\mathbf{Z}) \le \frac{1}{4}(\|\mathbf{X}\|_F^2+\|\mathbf{W}\|_F^2)(\|\mathbf{Y}\|_F^2+\|\mathbf{Z}\|_F^2)$

Bounding Π_1.

$$\begin{aligned}\Pi_1 \le& 2\|\mathbf{A}\|_2\left(\|\mathbf{A}\|_2 B^3+\|\mathbf{y}\|_2\right)\left(\|\mathbf{W}\|_F\|\mathbf{D}_U\|_F\|\mathbf{D}_V\|_F+\|\mathbf{V}\|_F\|\mathbf{D}_U\|_F\|\mathbf{D}_W\|_F+\|\mathbf{U}\|_F\|\mathbf{D}_V\|_F\|\mathbf{D}_W\|_F\right)\\ \le& 2\|\mathbf{A}\|_2\left(\|\mathbf{A}\|_2 B^3+\|\mathbf{y}\|_2\right)(3B\|\mathbf{D}\|_F^2)\\ =& 6\left(\|\mathbf{A}\|_2^2 B^4+\|\mathbf{A}\|_2\|\mathbf{y}\|_2 B\right)\|\mathbf{D}\|_F^2\end{aligned}$$

Bounding Π_2.

$$\Pi_2 \le \|\mathbf{A}\|_2^2\left[\|\mathbf{D}_U\circ\mathbf{V}\circ\mathbf{W}\|_F+\|\mathbf{U}\circ\mathbf{D}_V\circ\mathbf{W}\|_F+\|\mathbf{U}\circ\mathbf{V}\circ\mathbf{D}_W\|_F\right]^2 \le 9\|\mathbf{A}\|_2^2 B^4\|\mathbf{D}\|_F^2$$

Bounding Π_3.

$$\Pi_3 = \lambda(\|\mathbf{D}_U\|_F^2+\|\mathbf{D}_V\|_F^2+\|\mathbf{D}_W\|_F^2) \le \lambda\|\mathbf{D}\|_F^2$$

Bounding Π_4.

$$\Pi_4 \le \lambda(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2+\|\mathbf{W}\|_F^2)(\|\mathbf{D}_U\|_F^2+\|\mathbf{D}_V\|_F^2+\|\mathbf{D}_W\|_F^2) \le \lambda B^2\|\mathbf{D}\|_F^2$$

Bounding Π_5.

$$\begin{aligned}\Pi_5 \le& 4\lambda\Big[\frac{1}{4}(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2)(\|\mathbf{D}_U\|_F^2+\|\mathbf{D}_V\|_F^2)\\ &+\frac{1}{4}(\|\mathbf{U}\|_F^2+\|\mathbf{W}\|_F^2)(\|\mathbf{D}_U\|_F^2+\|\mathbf{D}_W\|_F^2)\\ &+\frac{1}{4}(\|\mathbf{V}\|_F^2+\|\mathbf{W}\|_F^2)(\|\mathbf{D}_V\|_F^2+\|\mathbf{D}_W\|_F^2)\Big]\\ \le& 2\lambda(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2+\|\mathbf{W}\|_F^2)(\|\mathbf{D}_U\|_F^2+\|\mathbf{D}_V\|_F^2+\|\mathbf{D}_W\|_F^2)\\ \le& 2\lambda B^2\|\mathbf{D}\|_F^2\end{aligned}$$

Combining the bounds from Π_1 to Π_5, we now can control the Lipschitz-continuity constant of the gradient function $\nabla f(\mathbf{U},\mathbf{V},\mathbf{W})$ within the following ball:

$$\mathcal{B}(\mathbf{0},B) := \{(\mathbf{U},\mathbf{V},\mathbf{W}) : \|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2+\|\mathbf{W}\|_F^2 \le B^2\}.$$

Note that all the iterations $\{(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)\}_{k\in\mathbb{N}}$ are living in this ball $\mathcal{B}(\mathbf{0},B)$ by Corollary 11.4.1, this Lipschitz-continuity constant also holds for ∇f evaluated along the solution path. □

11.4.3 Sufficient Decrease Property

Lemma 11.4.7 (Sufficient decrease property)**.** *Let the sequence* $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ *be generated by Algorithm 7. Then we have*

$$f(\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) - f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) \geq \lambda(\|\mathbf{U}_k - \mathbf{U}_{k-1}\|_F^2 + \|\mathbf{V}_k - \mathbf{V}_{k-1}\|_F^2 + \|\mathbf{W}_k - \mathbf{W}_{k-1}\|_F^2)$$

Proof. First recall the definitions of $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ in Algorithm 7:

$$\begin{aligned}
\mathbf{U}_k &= \arg\min_{\mathbf{U}} f(\mathbf{U}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1});\\
\mathbf{V}_k &= \arg\min_{\mathbf{V}} f(\mathbf{U}_k, \mathbf{V}, \mathbf{W}_{k-1});\\
\mathbf{W}_k &= \arg\min_{\mathbf{W}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W})
\end{aligned}$$

and the vanishing gradient equations in Lemma 11.4.2:

$$\begin{aligned}
&\nabla_{\mathbf{U}} f(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) = \mathbf{0},\\
&\nabla_{\mathbf{V}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_{k-1}) = \mathbf{0},\\
&\nabla_{\mathbf{W}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) = \mathbf{0}.
\end{aligned}$$

Now using the Taylor expansion, we expand f from $(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})$ to $(\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})$ and denote $\mathbf{D}_U := \mathbf{U}_{k-1} - \mathbf{U}_k$ and $\mathbf{U}(t) := t\mathbf{U}_{k-1} + (1-t)\mathbf{U}_k$. Then we have

$$\begin{aligned}
f(\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) =& f(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) + \langle \nabla_U f(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}), \mathbf{D}_U \rangle\\
&+ \frac{1}{2}\int_0^1 \left[\nabla_U^2 f(\mathbf{U}(t), \mathbf{V}_{k-1}, \mathbf{W}_{k-1})\right](\mathbf{D}_U, \mathbf{D}_U)\,\mathrm{d}\,t
\end{aligned}$$

Then in view of the diminishing gradient condition

$$\nabla_U f(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) = \mathbf{0}$$

and applying the λ-coordinate strong convexity of f by Lemma 11.4.1, we can further obtain that

$$f(\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) \geq f(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) + \lambda\|\mathbf{U}_{k-1} - \mathbf{U}_k\|_F^2$$

Using similar argument we can deduce that

$$\begin{aligned}
f(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) &\geq f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_{k-1}) + \lambda\|\mathbf{V}_{k-1} - \mathbf{V}_k\|_F^2\\
f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_{k-1}) &\geq f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) + \lambda\|\mathbf{W}_{k-1} - \mathbf{W}_k\|_F^2
\end{aligned}$$

The proof then can be completed by noting that:

$$
\begin{aligned}
& f(\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) - f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_{k-1}) \\
=& f(\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) - f(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1} \\
& + f(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1} - f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_{k-1}) \\
& + f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_{k-1}) - f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_{k-1}).
\end{aligned}
$$

□

11.5 Convergence of Algorithm 7

Theorem 11.5.1 (Subsequence convergence)**.** *Let the iterates sequence $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ be generated by Algorithm 7, the following holds*

1. *The sequence of the function values $\{f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ is non-increasing and it is convergent to some finite value $\lim_{k\to\infty} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) = \overline{f}$ for some unique $\overline{f} \geq 0$.*

2. *The iterates sequence $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ is regular, i.e., the difference between iterates sequence is convergent, i.e.*

$$
\begin{aligned}
\lim_{k\to\infty} \|\mathbf{U}_{k+1} - \mathbf{U}_k\|_F = 0, \\
\lim_{k\to\infty} \|\mathbf{V}_{k+1} - \mathbf{V}_k\|_F = 0, \\
\lim_{k\to\infty} \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F = 0.
\end{aligned}
$$

3. *Denote $\mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$ as the set of all limit points of those convergent subsequences of $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ (which depends on the initialization $(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$). Then all limit points have the same function value*

$$
f(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}}) = \overline{f}, \ \ \forall (\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}}) \in \mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0).
$$

4. *The gradient at each iterate is bounded: for all $k \geq 1$,*

$$
\|\nabla f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F \leq \sqrt{2}\mathcal{L}_g \|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})\|_F. \tag{11.10}
$$

   This implies that each $(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}}) \in \mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$ is a critical point of f, i.e. $\nabla f(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}}) = \mathbf{0}$. Further, $\mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$ is a nonempty, compact and connect set, satisfying

$$
\lim_{k\to\infty} \operatorname{dist}((\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k), \mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)) = 0.
$$

Proof. Now we prove Theorem 11.5.1.

1. The first part is because the sequence of the function values $\{f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ is non-increasing by Lemma 11.4.7 and lower-bounded, and hence convergent.

2. To prove the second part, we denote

$$\Delta_k^2 := \|\mathbf{U}_{k+1} - \mathbf{U}_k\|_F^2 + \|\mathbf{V}_{k+1} - \mathbf{V}_k\|_F^2 + \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F^2$$

and recursively use sufficient decreasing property Lemma 11.4.7:

$$f(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0) - f(\mathbf{U}_N, \mathbf{V}_N, \mathbf{W}_N) \geq \lambda \sum_{k=0}^{N} \Delta_k^2,$$

which then implies that (using $f \geq 0$)

$$\sum_{k=0}^{N} \Delta_k^2 \leq \frac{f(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)}{\lambda}, \quad \forall N \in \mathbb{N}.$$

Therefore, we identify that the sequence $\{\sum_{k=0}^{N} \Delta_k^2\}_{N\in\mathbb{N}}$ is upper bounded and non-decreasing, hence convergent. We are therefore guaranteed that $\lim_{k\to\infty} \Delta_k^2 = 0$. By definition of Δ_k^2, we finally get that $\lim_{k\to\infty} \|\mathbf{U}_{k+1} - \mathbf{U}_k\|_F = 0, \lim_{k\to\infty} \|\mathbf{V}_{k+1} - \mathbf{V}_k\|_F = 0, \lim_{k\to\infty} \|\mathbf{W}_{k+1} - \mathbf{W}_k\|_F = 0$.

3. For the third part, by the boundedness of the sequence $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$, we extract an arbitrary convergent subsequence $\{(\mathbf{U}_{k_m}, \mathbf{V}_{k_m}, \mathbf{W}_{k_m})\}_{m\in\mathbb{N}}$ with the limit $(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}})$. Then take the limit on subsequence:

$$\lim_{m\to\infty} f(\mathbf{U}_{k_m}, \mathbf{V}_{k_m}, \mathbf{W}_{k_m}) = f(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}}),$$

where we have used the continuity of the objective function f. The proof of this part is completed by noting that the function-value sequence $\{f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ is convergent and hence all its subsequence must be also convergent and converge to the same limiting point.

4. For this part, we first show that

$$\lim_{k\to\infty} \nabla f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) = \lim_{k\to\infty} \begin{bmatrix} \nabla_{\mathbf{U}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) \\ \nabla_{\mathbf{V}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) \\ \nabla_{\mathbf{W}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) \end{bmatrix}$$

vanishes. Towards that, we use the first-order optimality condition (11.8) to get that for $k \geq 1$,

$$\begin{aligned} &\nabla_{\mathbf{U}} f(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1}) = \mathbf{0}; \\ &\nabla_{\mathbf{V}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_{k-1}) = \mathbf{0}; \\ &\nabla_{\mathbf{W}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) = \mathbf{0}. \end{aligned}$$

This together with the Lipschitz-continuity of gradient ∇f Lemma 11.4.5 implies that

$$\begin{aligned} \|\nabla_{\mathbf{U}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F &= \|\nabla_{\mathbf{U}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - \nabla_{\mathbf{U}} f(\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})\|_F \\ &\leq \mathcal{L}_g \|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\mathbf{U}_k, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})\|_F; \end{aligned}$$

and

$$\begin{aligned}\|\nabla_{\mathbf{V}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F &= \|\nabla_{\mathbf{V}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - \nabla_{\mathbf{V}} f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_{k-1})\|_F \\ &\leq \mathcal{L}_g \|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_{k-1})\|_F.\end{aligned}$$

Combining the above two and denote $\mathbf{D}_k := (\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})$, we obtain that

$$\|\nabla f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F \leq \sqrt{2}\mathcal{L}_g \|\mathbf{D}_k\|_F.$$

Then let k go to infinity, we get $\|\nabla f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F$ converge 0 by Part (ii) of Theorem 11.5.1. Now, we extract a convergent subsequence $\{(\mathbf{U}_{k_m}, \mathbf{V}_{k_m}, \mathbf{W}_{k_m})\}_{m\in\mathbb{N}}$ with limit $(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}})$ and use the continuity of ∇f (hence $\|\nabla f\|_F$) to get

$$\lim_{m\to\infty} \|\nabla f(\mathbf{U}_{k_m}, \mathbf{V}_{k_m}, \mathbf{W}_{k_m})\|_F = \|\nabla f(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}})\|_F.$$

Hence any limit point $(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}}) \in \mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$ is a critical point of f. The remaining proof in this part follows from a similar argument as in [213, Lemma 5(iii)] by identifying that $\{f(\mathbf{U}_{k_m}, \mathbf{V}_{k_m}, \mathbf{W}_{k_m})\}_{m\in\mathbb{N}}$ is bounded (Lemma 11.4.4) and regular (Part (ii) of Theorem 11.5.1).

□

Finally, combining with the Kurdyka-Łojasiewicz property [212,213,239,241] for characterization of the geometry of objective function around its critical points, we can obtain a stronger convergence result than Theorem 11.5.1. We put the proof in Section H.1 in the supplement.

Theorem 11.5.2 (Sequence convergence)**.** *The iterates sequence $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ generated by Algorithm 7 is convergent to its unique limit point $(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}})$, which is a critical point of the objective function f* (11.7)*. Moreover, the convergence rate of $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ is at least sub-linear.*

Remark.

- Theorem 11.5.2 is much stronger than Theorem 11.5.1, since we are not guaranteed that the iterates $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ generated by Algorithm 7 would converge to a limit point only by Theorem 11.5.1. Theorem 11.5.2 fulfills this gap by directly showing the iterates $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ generated by Algorithm 7 converge to a critical point, and as an consequence, the set of limit point $\mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$ becomes a singleton.
- The main part of the proof is based on the Kurdyka-Łojasiewicz property [212,213,239,241] for characterization of the geometry of objective function (including its constraints) around its critical points, which plays a key role in our sequel analysis.

11.6 Extension to Constrained Minimization

Recall Algorithm 8 aims to solve the "factored" tensor nuclear norm regularized minimization problem (11.7), where the regularization parameter λ should be chosen with respect to the observation noise level. However, in the noise-free observation regime, as long as λ is positive, we remark that the estimator $(\mathbf{U}^\lambda, \mathbf{V}^\lambda, \mathbf{W}^\lambda)$ returned by Algorithm 8 is always biased, due to the non-negativity of the regularization term. In practical applications, this is fine due to the inevitable observation noise in the real data. For theoretical purpose, it is of ultimate interest to study the constrained minimization

$$\begin{aligned} &\underset{\mathbf{U},\mathbf{V},\mathbf{W}}{\text{minimize}} \Big[(\|\mathbf{U}\|_F^2 + \mathrm{tr}(\mathbf{V}^\top\mathbf{V} \odot \mathbf{W}^\top\mathbf{W})) + (\|\mathbf{V}\|_F^2 + \mathrm{tr}(\mathbf{U}^\top\mathbf{U} \odot \mathbf{W}^\top\mathbf{W})) + (\|\mathbf{W}\|_F^2 + \mathrm{tr}(\mathbf{U}^\top\mathbf{U} \odot \mathbf{V}^\top\mathbf{V}))\Big] \\ &\text{subject to} \quad \mathbf{y} = \mathrm{A}(\mathbf{U} \circ \mathbf{V} \circ \mathbf{W}) \end{aligned} \tag{11.11}$$

Therefore, it is necessary to extend Algorithm 8 to solve the constrained optimization (11.11).

Algorithm 8 AltMinC

1: **Initialization:** λ, $\beta \in (0,1)$, $k = 0$, and $(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$.
2: **while** stop criterion not meet **do**
3: $\quad (\mathbf{U}_{k+1}, \mathbf{V}_{k+1}, \mathbf{W}_{k+1}) = \mathrm{AlgMin}^\lambda(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)$
4: $\quad k = k + 1$.
5: $\quad \lambda = \beta\lambda$.
6: **end while**
7: **Output:** factorization $(\mathbf{U}^k, \mathbf{V}^k, \mathbf{W}^k)$.

Remark. By Theorem 11.5.2, we have shown the global sequence convergence of Algorithm 7 for any initialization $(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$. Note that we can view Algorithm 8 as a sequence of Algorithm 7 with a series of decreasing λ, then a direct consequence of Theorem 11.5.2 gives the convergence of Algorithm 8.

11.7 Experiments on Synthetic and Image Data

In this section, we conduct experiments on both synthetic data and real data to illustrate the performance of our proposed algorithms and compare it to other state-of-the-art ones, e.g., **LRTC** [242], **HaLRTC** [243], and **FaLRTC** [243], concerning both synthetic tensor recovery and real image estimation performance.

11.7.1 Experiments on Synthetic Data

In this subsection, we mainly test Algorithm 7 on synthetic data. Particularly, we first generate a rank-3 ground-true tensor $\boldsymbol{\mathcal{T}}^\star \in \mathbb{R}^{40\times20\times10}(i.e., n_1 = 40, n_2 = 20, n_3 = 10, r = 3)$ by $\boldsymbol{\mathcal{T}}^\star = \mathbf{U}^\star \circ \mathbf{V}^\star \circ \mathbf{W}^\star$ with $\mathbf{U}^\star \in \mathbb{R}^{n_1\times r}, \mathbf{V}^\star \in \mathbb{R}^{n_2\times r}, \mathbf{W}^\star \in \mathbb{R}^{n_3\times r}$ being three random matrices. However, we are only allowed to observe a small partial of $\boldsymbol{\mathcal{T}}^\star$, with the observed positions denoted by Ω. That is, we only have the knowledge of $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{40\times20\times10}$ with its most

entries being zero: $\boldsymbol{\mathcal{Y}}_{i,j,k} = \begin{cases} \boldsymbol{\mathcal{T}}^\star_{i,j,k} & (i,j,k) \in \Omega \\ 0 & (i,j,k) \notin \Omega \end{cases}$, which is a tensor completion problem. Here we are particularly interested in the situation that $m \ll n_1 n_2 n_3$ with m denoting the cardinality of Ω. Theoretically, using tensor nuclear norm $\|\boldsymbol{\mathcal{T}}\|_*$

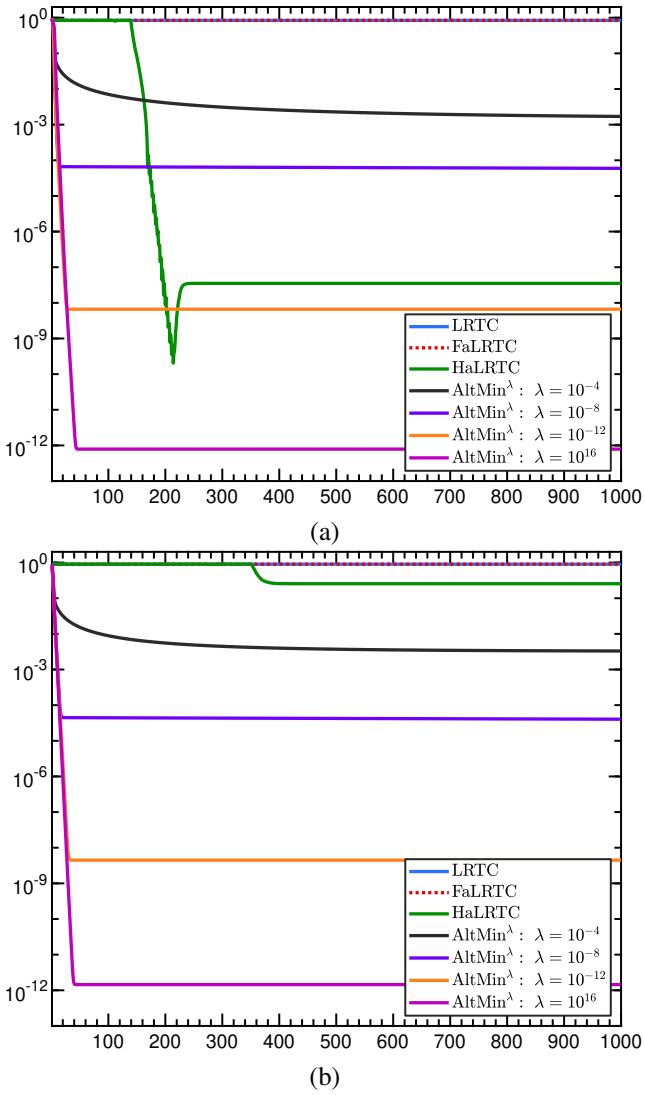
Note that the low-rank tensor completion problem is covered by our general linear model eq. (11.3) (c.f. Example 11.2.2). Therefore, we can apply the proposed Algorithm 7: AltMin^λ to recover $\boldsymbol{\mathcal{T}}^\star$. We remark that traditionally the parameter λ in AltMin^λ in Algorithm 7 is set to as the noise level. When the observation is noise-free, we can set λ as a very small constant[41]. Now we apply the proposed Algorithm 7 for different values of λ and several state-of-the-art tensor completion algorithms (e.g., **LRTC** [242], **HaLRTC** [243], and **FaLRTC** [243] to recover the tensor $\boldsymbol{\mathcal{T}}^\star$. Moreover, to test the robustness of these algorithms, we will apply them under two different data-missing ratios $\frac{m}{n_1 n_2 n_3}$ and we record their relative error $\|\widehat{\boldsymbol{\mathcal{T}}}(k) - \boldsymbol{\mathcal{T}}^\star\|_F / \|\boldsymbol{\mathcal{T}}^\star\|_F$ with $\widehat{\boldsymbol{\mathcal{T}}}(k)$ denoting the recovered tensor for a certain algorithm after kth iterations. Figure 11.1 shows the proposed Algorithm 7 (even for different λ) achieves better recovery performance than other algorithms in both low- and large-missing ratios, implying both the superiority and robustness of Algorithm 7 in recovering missing data.

11.7.2 Experiments on Real Image

Given any original color image (a tensor), say $\boldsymbol{\mathcal{T}}^\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ (note for a color image $n_3 = 3$), assume we have only access to a smaller number of random distributed positions of the original color image $\boldsymbol{\mathcal{T}}^\star$, i.e., we have access to the $\boldsymbol{\mathcal{Y}}$ satisfying $\boldsymbol{\mathcal{Y}}_\Omega = \boldsymbol{\mathcal{T}}^\star_\Omega$ and $\boldsymbol{\mathcal{Y}}_{\Omega^c} = \mathbf{0}$ where Ω^c denoting the complement set of Ω. We assume the total number of observations $m \ll n_1 n_2 n_3$.

Note since in the situation of real-data images the rank of $\boldsymbol{\mathcal{T}}^\star$ is unknown (and NP-hard to compute), we therefore can pick an appropriate r for Algorithm 7 to recover the ground-truth color image $\boldsymbol{\mathcal{T}}^\star$. Remarkably, due to the Burer-Monteiro factorization structure embedded in Algorithm 7, we calm that the rank of the recovered color image by Algorithm 7 with a specific r, denoted as $\widehat{\boldsymbol{\mathcal{T}}}_r$, is alway satisfying $\mathrm{rank}(\widehat{\boldsymbol{\mathcal{T}}}_r) \leq r$. That is, the proposed Algorithm 7 can guarantee the low-rankness of the recovered tensors from the true definition of tensor rank, which is what those state-of-the-art algorithms (that tries to promote the low-rankness of recovered tensors only using matrix nuclear norms (c.f. [242–244])). To well display the recovery performance, we also record the relative recovery error as a function of the iteration number k: $\frac{\|\widehat{\boldsymbol{\mathcal{T}}}(k) - \boldsymbol{\mathcal{T}}^\star\|_F}{\|\boldsymbol{\mathcal{T}}^\star\|_F}$ with $\widehat{\boldsymbol{\mathcal{T}}}(k)$ denoting the recovered tensor by those different tensor completion algorithms after kth iterations. The main algorithms that we will compare with are several powerful state-of-the-art algorithms in using tensor completion for color image recovery, e.g., **LRTC** [242], **HaLRTC** [243], and **FaLRTC** [243]. All the parameters of these algorithms are set up according to their suggested values [242, 243]. The comparison results are summarized in Figure 11.2, from which we can see that

[41] Even we can just set $\lambda = 0$ so that Algorithm 7 reduces to the traditional alternating least squares [76].



(a)

(b)

Figure 11.1: Performing Algorithm 7, **LRTC**, **HaLRTC**, and **FaLRTC** to recover $\boldsymbol{\mathcal{T}}^{\star}$ for two different missing-data ratio and recording their relative recovery errors $\|\widehat{\boldsymbol{\mathcal{T}}}(k) - \boldsymbol{\mathcal{T}}^{\star}\|_F / \|\boldsymbol{\mathcal{T}}^{\star}\|_F$ versus iteration, where $\widehat{\boldsymbol{\mathcal{T}}}(k)$ denotes the recovered tensor by certain algorithm after k-th iteration. (a) missing-data ratio=70% and (b) missing-data ratio=80%.

- Algorithm 7 converges much faster, as it requires much fewer iterations for Algorithm 7 to converge; This is in sharp contrast to other algorithms, which instead need several thousands of iterations to converge.
- Algorithm 7 converges to a “better" solution compared with other algorithms by comparing their relative recovery errors.

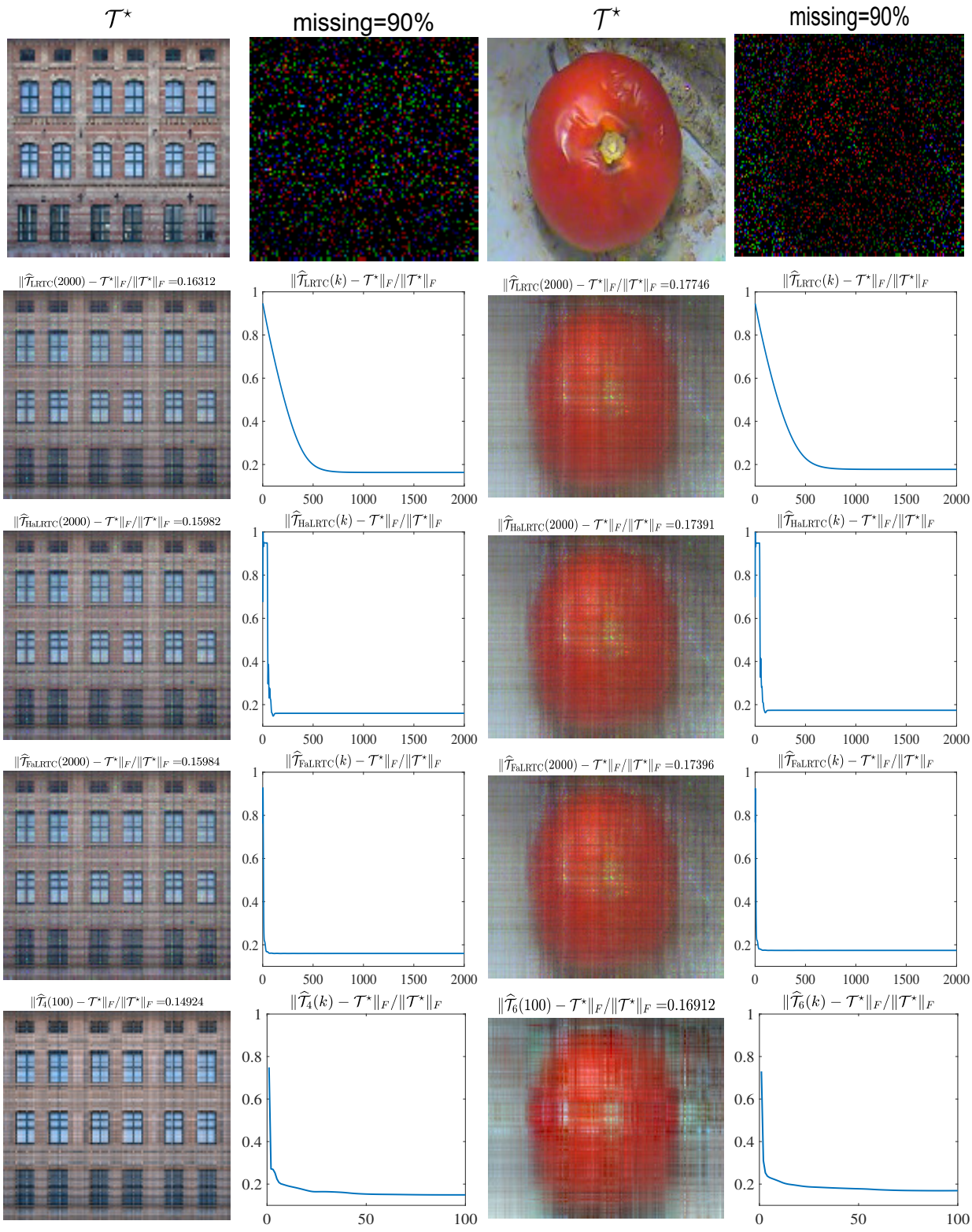


Figure 11.2: Compare Algorithm 7 with **LRTC**, **HaLRTC**, and **FaLRTC** in missing image recovery in term of the relative recovery errors versus iteration. (**Left**) Test on the House image; (**Right**) Test on the Tomato image. Here we denote the recovered image by Algorithm 7 by $\widehat{\boldsymbol{\mathcal{T}}}_r$ with r indicating the input rank of the algorithm. Both show that the proposed Algorithm 7 converges with fewer iterations and to a better solution in term of the relative recovery errors.

CHAPTER 12

SPHERICAL CLUSTERING VIA ALTERNATING MINIMIZATION

Principal Component Analysis (PCA) is one of the most important methods to handle high dimensional data. However, most of the studies on PCA aim to minimize the loss after projection, which usually measure the Euclidean distance, though in some fields, angle distance is known to be more important and critical for analysis. In this work[42], we propose a method by adding constraints on factors to unify the Euclidean distance and angle distance. However, due to the nonconvexity of the objective and constraints, the optimized solution is not easy to obtain. We propose an alternating linearized minimization method to solve it with provable convergence rate and guarantee. Experiments on synthetic data and real-world datasets have validated the effectiveness of our method and demonstrated its advantages over state-of-art clustering methods.

12.1 Introduction

In many real-world applications such as text categorization and face recognition, the dimensions of data are usually very high. Dealing with high-dimensional data is computationally expensive while noise or outliers in the data can increase dramatically as the dimension increases. Dimension reduction is one of the most important and effective methods to handle high dimensional data [245–247]. Among the dimension reduction methods, Principal Component Analysis (PCA) is one of the most widely used methods due to its simplicity and effectiveness.

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of correlated variables into a set of linearly uncorrelated principal directions. Usually the number of principal directions is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal direction has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding direction has the highest variance under the constraint that it is orthogonal to the preceding directions. The resulting vectors are an uncorrelated orthogonal basis set.

When data points lie in a low-dimensional manifold and the manifold is linear or nearly-linear, the low-dimensional structure of data can be effectively captured by a linear subspace spanned by the principal PCA directions.

More specifically, let $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_n] \in \mathbb{R}^{m\times n}$ be n data points in m-dimensional space while $\mathbf{U} = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_r] \in \mathbb{R}^{m\times r}$ contains the principal directions and $\mathbf{V} = [\mathbf{v}_1 \ \ldots \ \mathbf{v}_n] \in \mathbb{R}^{r\times n}$ contains the principal components (data projects along the principal directions). Generally speaking, there can be two formulations for PCA:

- Covariance-based approach, which computes the covariance matrix $\mathbf{C} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top = \mathbf{X}\mathbf{X}^\top$. Here we assume the data are already centered, i.e., $\bar{\mathbf{x}} = 0$, and we drop the factor $\frac{1}{n-1}$ which does not affect $\mathbf{U}$.

[42] This is a joint work with Kai Liu, Hua Wang and Gongguo Tang [12].

The principal directions are obtained as:

$$\underset{\mathbf{U}^\top\mathbf{U}=\mathbf{I}}{\text{maximize}}\, \text{tr}(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}). \tag{12.1}$$

- Matrix low-rank approximation-based approach. Let $\mathbf{X} \approx \mathbf{U}\mathbf{V}$, we solve:

$$\underset{\mathbf{U}^\top\mathbf{U}=\mathbf{I}}{\text{minimize}}\, \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2 = \sum_{i,j}[\mathbf{X}_{ij} - (\mathbf{U}\mathbf{V})_{ij}]^2. \tag{12.2}$$

Taking the derivative w.r.t. $\mathbf{V}$ and setting it to zero, we have $\mathbf{V} = \mathbf{X}^\top\mathbf{U}$, and Eq. (12.2) reduces to Eq. (12.1). Therefore, the solutions to these two approaches are identical. In this work, we mainly focus on the second formulation.

12.2 Motivation

In Eq. (12.2), the objective function measures the gap between original data $\mathbf{X}$ and approximation after projection $\mathbf{U}\mathbf{V}$, which is based on squared Euclidean distance measurements and treat each feature as equally important. However, in the real world, there are some given datasets which are preprocessed to be normalized and different features may have various significance. Thus distance-based measurement method may yield poor results. On the other side, similarity-based measurement methods such as angle distance have been proved to be more efficient in some applications, including information retrieval [248], signal processing [249], metric learning [250], etc.. Though one can calculate the similarity after projection, still this appears to be more or less awkward and inefficient. Thus, deriving some methods which can directly measure angle distance from PCA is vitally important. However, to our best knowledge, it has not been studied yet.

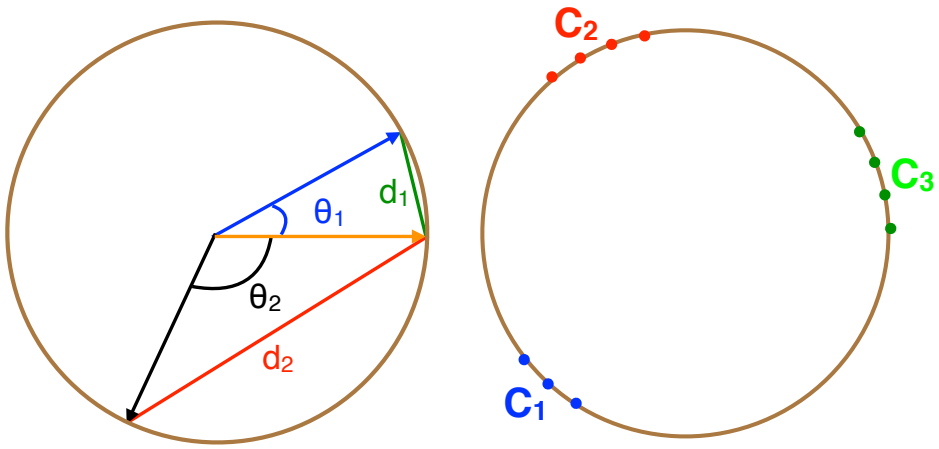


Figure 12.1: Larger angles ($\theta_2 > \theta_1$) in the sphere will have larger Euclidean distance, and vice versa, which unifies the cosine similarity and Euclidean distance simultaneously.

Motivated by the above observations and a previous work [251], in this work we propose a spherical-PCA model which can unify the Euclidean distance and angle distance. By noticing that larger angle in the sphere in Figure 12.1

also has larger Euclidean distance, we can add the normalization constraint to the component matrix, where the norm of each column in $\mathbf{V}$ is 1 to guarantee the spherical distribution of components:

$$\underset{\mathbf{U}\in\mathbb{R}^{m\times r},\mathbf{V}\in\mathbb{R}^{r\times n}}{\text{minimize}} \|\mathbf{X}-\mathbf{U}\mathbf{V}\|_F^2 = \sum_{i,j}[\mathbf{X}_{ij}-(\mathbf{U}\mathbf{V})_{ij}]^2 \quad s.t. \quad \mathbf{U}\in\mathbb{U},\ \mathbf{V}\in\mathbb{V} \tag{12.3}$$

where we define:

$$\begin{aligned} \mathbb{U} &:= \{\mathbf{U} : \mathbf{U}^\top\mathbf{U} = \mathbf{I}\}, \\ \mathbb{V} &:= \{\mathbf{V} : \|\mathbf{v}_j\| = 1\ \forall j\}, \end{aligned} \tag{12.4}$$

where $\|\cdot\|$ denotes ℓ_2 norm for vectors and denotes the spectral norm for matrices. Suppose the component is spherically distributed, then the Euclidean distance between $\mathbf{v}_i$ and $\mathbf{v}_j$ is:

$$\begin{aligned} \|\mathbf{v}_i-\mathbf{v}_j\|^2 &= \|\mathbf{v}_i\|^2+\|\mathbf{v}_j\|^2-2\langle\mathbf{v}_i,\mathbf{v}_j\rangle \\ &= \|\mathbf{v}_i\|^2+\|\mathbf{v}_j\|^2-2\frac{\langle\mathbf{v}_i,\mathbf{v}_j\rangle}{\|\mathbf{v}_i\|\|\mathbf{v}_j\|} \\ &= 2-2\cos(\theta),\ \theta\in[0,\pi] \end{aligned} \tag{12.5}$$

which is equivalent to angle distance that bigger angle θ will result in larger Euclidean distance, and vice versa.

Remark 12.2.1. In traditional PCA, without the normalization constraint on each column of $\mathbf{v}$, the optimized solution to Eq. (12.2) can barely satisfy the spherical distribution. Since r is usually less than m, PCA will lose some component more or less, thus $\mathbf{x}_i \neq \mathbf{U}\mathbf{v}_i$ and usually $\|\mathbf{x}_i\| \neq \|\mathbf{U}\mathbf{v}_i\|$ (they may be equal, but it barely happens) . We have $\|\mathbf{x}_i\|^2 = 1$ for normalized data and if $\|\mathbf{v}_i\|^2 = 1$ then $\|\mathbf{U}\mathbf{v}_i\|^2 = \mathrm{tr}(\mathbf{v}_i^\top\mathbf{U}^\top\mathbf{U}\mathbf{v}_i) = \mathrm{tr}(\mathbf{v}_i^\top\mathbf{v}_i) = \|\mathbf{v}_i\|^2 = 1$, which leads a contradiction, thus the constraint on $\mathbf{V}$ is necessary to guarantee our motivation.

12.3 Formulation And Algorithm

12.3.1 Objective Function with Proximal Term

We first denote:

$$h(\mathbf{U},\mathbf{V}) = \|\mathbf{X}-\mathbf{U}\mathbf{V}\|_F^2 = \sum_{j=1}^{n}\|\mathbf{x}_j-\mathbf{U}\mathbf{v}_j\|^2 \quad s.t. \quad \mathbf{U}\in\mathbb{U},\mathbf{V}\in\mathbb{V} \tag{12.6}$$

By noting the nonconvexity of Eq. (12.3), where no closed solution exists, we propose an alternating minimization method to get the optimized solution as: given kth iterate of $\mathbf{V}$ varaible $\mathbf{V}(k) = [\mathbf{v}_1(k)\ \cdots\ \mathbf{v}_n(k)]$,

$$\begin{aligned} \mathbf{U}(k+1) &= \underset{\mathbf{U}\in\mathbb{U}}{\arg\min}\|\mathbf{X}-\mathbf{U}\mathbf{V}(k)\|_F^2; \\ \mathbf{v}_j(k+1) &= \underset{\|\mathbf{v}\|=1}{\arg\min}\|\mathbf{x}_j-\mathbf{U}(k+1)\mathbf{v}\|_2^2,\ \forall j \end{aligned} \tag{12.7}$$

Note that when the constraints $\mathbf{U}\in\mathbb{U},\mathbf{V}\in\mathbb{V}$, the problem (12.6) is known as the nonconvex matrix factorization problems, which have been well-studied [6, 8]. This work focus on develop efficient and provable algorithm to deal with (12.6) with the constraints $\mathbf{U}\in\mathbb{U},\mathbf{V}\in\mathbb{V}$. Note that the proximal algorithm recently has been successfully applied to a wide variety of situations: convex optimization, nonmonotone operators [252, 253] with various applications to nonconvex programming. It was first introduced by Rockafellar [254] as an approximation regularization

method in convex optimization and in the study of variational inequalities associated to maximal monotone operators.

Considering the fact that the objective function in Eq. (12.3) is nonconvex w.r.t. $\mathbf{U}$ and $\mathbf{V}$, and the constraint on $\mathbf{U}$ and $\mathbf{V}$ are also nonconvex, we consider adding proximal term and optimize the solution as: with the alternating linearized minimization solutions becomes:

$$\begin{aligned}\mathbf{U}(k+1) &= \arg\min_{\mathbf{U}\in\mathbb{U}} \langle \mathbf{U}-\mathbf{U}(k), \nabla_{\mathbf{U}} h(\mathbf{U}(k),\mathbf{V}(k))\rangle + \frac{\mu}{2}\|\mathbf{U}-\mathbf{U}(k)\|_F^2; \\ \mathbf{v}_j(k+1) &= \arg\min_{\|\mathbf{v}\|=1} \langle \mathbf{v}-\mathbf{v}_j(k), \nabla_{\mathbf{v}_j} h(\mathbf{U}(k+1),\mathbf{V}(k))\rangle + \frac{\lambda}{2}\|\mathbf{v}-\mathbf{v}_j(k)\|^2,\ \forall j\end{aligned} \tag{12.8}$$

Remark 12.3.1. We add the proximal term to make the new updating solution will not be too far from the previous step to avoid drastic changes. One can see that when the proximal term regularization parameters μ, λ are sufficiently large, they will dominate the objective function. Moreover, we can take the linearized minimization as to minimize the objective with Taylor expansion by making use of first order (linear) information.

12.3.2 Proposed Algorithm

Given the alternating minimization objective in Eq. (12.8), now we turn to provide detailed (closed) updating algorithm.

We first derive the solution for $\mathbf{U}$ and before that we give a useful lemma that is similar to [255, Theorem 1] and [256, Theorem 1]:

Lemma 12.3.1. $\text{maximize}_{\mathbf{U}^\top\mathbf{U}=\mathbf{I}} \operatorname{tr}(\mathbf{U}^\top\mathbf{M})$ *is given by* $\mathbf{U}=\mathbf{A}\mathbf{B}^\top$, *where* $[\mathbf{A},\mathbf{\Sigma},\mathbf{B}] = \operatorname{svd}(\mathbf{M})$.

Proof. On one hand, we have:

$$\operatorname{tr}(\mathbf{U}^\top\mathbf{M}) = \operatorname{tr}(\mathbf{U}^\top\mathbf{A}\mathbf{\Sigma}\mathbf{B}^\top) = \operatorname{tr}(\mathbf{P}\mathbf{\Sigma}), \tag{12.9}$$

where $\mathbf{P}=\mathbf{B}^\top\mathbf{U}^\top\mathbf{A}$ is an orthogonal matrix since

$$\mathbf{P}\mathbf{P}^\top = (\mathbf{B}^\top\mathbf{U}^\top\mathbf{A})(\mathbf{B}^\top\mathbf{U}^\top\mathbf{A})^\top = \mathbf{I}.$$

Thus every element including the diagonal of $\mathbf{P}$ is no larger than 1. Then we have:

$$\operatorname{tr}(\mathbf{P}\mathbf{\Sigma}) \leq \operatorname{tr}(\mathbf{\Sigma}) \tag{12.10}$$

On the other hand, when $\mathbf{U}=\mathbf{A}\mathbf{B}^\top$, we have

$$\operatorname{tr}(\mathbf{U}^\top\mathbf{M}) = \operatorname{tr}(\mathbf{B}\mathbf{A}^\top\mathbf{A}\mathbf{\Sigma}\mathbf{B}^\top) = \mathbf{\Sigma}.$$

Thus $\mathbf{U}=\mathbf{A}\mathbf{B}^\top$ is the optimized solution to maximize the objective. □

Accordingly, we have:

$$\begin{aligned}\mathbf{U}(k+1) &= \arg\min_{\mathbf{U}^\top\mathbf{U}=\mathbf{I}} \langle \mathbf{U}-\mathbf{U}(k), \nabla_{\mathbf{U}} h(\mathbf{U}(k),\mathbf{V}(k))\rangle + \frac{\mu}{2}\|\mathbf{U}-\mathbf{U}(k)\|_F^2 \\ &= \arg\max_{\mathbf{U}^\top\mathbf{U}=\mathbf{I}} \langle \mathbf{U}, \mathbf{M}(k)\rangle = \mathbf{Y}\mathbf{Z}^\top\end{aligned} \tag{12.11}$$

Algorithm 9 Alternating Linearized Minimization for Problem Eq. (12.6)

Input: data $\mathbf{X} \in \mathbb{R}^{m\times n}$, rank of factors r, regularization parameters λ, μ, number of iterations K
Initialization: $\mathbf{U}(0) \in \mathbb{R}^{m\times r}, \mathbf{V}(0) \in \mathbb{R}^{r\times n}$
for $k = 1, \ldots, K$
  Optimize $\mathbf{U}(k)$ via Eq. (12.11)
  Optimize $\mathbf{V}(k)$ via Eq. (12.12) for $j = 1, \ldots, n$
end
Output: $\mathbf{U}(K)$ and $\mathbf{V}(K)$

where $\mathbf{M}(k) := 2(\mathbf{X} - \mathbf{U}(k)\mathbf{V}(k))\mathbf{V}(k)^\top + \mu\mathbf{U}(k)$ and $\mathbf{Y}, \mathbf{Z}$ is obtained from $[\mathbf{Y}, \mathbf{\Sigma}, \mathbf{Z}] = \mathrm{svd}(\mathbf{M}(k))$.

Then we compute $\mathbf{V}(k+1)$:

$$\begin{aligned}\mathbf{v}_j(k+1) &= \arg\min_{\|\mathbf{v}\|=1} \langle \mathbf{v} - \mathbf{v}_j(k), \nabla_{\mathbf{v}_j} h(\mathbf{U}(k+1), \mathbf{V}(k))\rangle + \frac{\lambda}{2}\|\mathbf{v} - \mathbf{v}_j(k)\|^2 \\ &= \arg\max_{\|\mathbf{v}\|=1} \langle \mathbf{v}, \mathbf{q}\rangle \\ &= \frac{\mathbf{q}}{\|\mathbf{q}\|}, \quad \text{for } j = 1, \ldots, n,\end{aligned} \tag{12.12}$$

where $\mathbf{q} := 2\mathbf{U}(k+1)^\top \mathbf{x}_j + (\lambda - 2)\mathbf{v}_j(k)$.

12.4 Convergence Analysis

In the following case, we let $\mathbb{U}$ and $\mathbb{V}$ be as defined in Eq. (12.4), and show the convergence of our proposed algorithm in the last section.

To begin with, we first show that $h(\mathbf{U}, \mathbf{V})$ has Lipschitz continuous gradient at $\mathbf{U} \in \mathbb{U}, \mathbf{V} \in \mathbb{V}$, which will be very useful for the following convergence analysis.

Proposition 12.4.1. *$h(\mathbf{U}, \mathbf{V})$ has Lipschitz continuous gradient at $\mathbf{U} \in \mathbb{U}, \mathbf{V} \in \mathbb{V}$, where $\mathbb{U}$ and $\mathbb{V}$ are defined in Eq.* (12.4)*. That is, there exists a constant L_c such that*

$$\|\nabla h(\mathbf{U}, \mathbf{V}) - \nabla h(\mathbf{U}', \mathbf{V}')\|_F \le L_c \|(\mathbf{U}, \mathbf{V}) - (\mathbf{U}', \mathbf{V}')\|_F \tag{12.13}$$

for all $\mathbf{U}, \mathbf{U}' \in \mathbb{U}$ and $\mathbf{V}, \mathbf{V}' \in \mathbb{V}$. Here $L_c > 0$ is referred to as the Lipschitz constant.

Proof of Proposition 12.4.1. It is equivalent to show $\|\nabla^2 h(\mathbf{U}, \mathbf{V})\| \le L_c$ for all $\mathbf{U} \in \mathbb{U}, \mathbf{V} \in \mathbb{V}$. Standard computations give the Hessian quadrature form $[\nabla^2 h(\mathbf{U}, \mathbf{V})](\mathbf{\Delta}, \mathbf{\Delta})$ for any $\mathbf{\Delta} = \begin{bmatrix} \mathbf{\Delta}_\mathbf{U} \\ \mathbf{\Delta}_\mathbf{V}^\top \end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$ (where $\mathbf{\Delta}_\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{\Delta}_\mathbf{V} \in \mathbb{R}^{r\times n}$) as

$$[\nabla^2 h(\mathbf{U}, \mathbf{V})](\mathbf{\Delta}, \mathbf{\Delta}) = \|\mathbf{\Delta}_\mathbf{U}\mathbf{V} + \mathbf{U}\mathbf{\Delta}_\mathbf{V}\|_F^2 + 2\langle \mathbf{U}\mathbf{V} - \mathbf{X}, \mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}\rangle \tag{12.14}$$

which gives:

$$\begin{aligned}\|\nabla^2 h(\mathbf{U},\mathbf{V})\| &= \underset{\|\boldsymbol{\Delta}\|_F=1}{\text{maximize}} \left|[\nabla^2 h(\mathbf{U},\mathbf{V})](\boldsymbol{\Delta},\boldsymbol{\Delta})\right| \\ &\leq \underset{\|\boldsymbol{\Delta}\|_F=1}{\text{maximize}} \|\boldsymbol{\Delta}_\mathbf{U}\mathbf{V}+\mathbf{U}\boldsymbol{\Delta}_\mathbf{V}\|_F^2 + 2\left|\langle \mathbf{U}\mathbf{V}-\mathbf{X}, \boldsymbol{\Delta}_\mathbf{U}\boldsymbol{\Delta}_\mathbf{V}\rangle\right| \\ &\leq 2(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2+\|\mathbf{U}\|_F\|\mathbf{V}\|_F+\|\mathbf{X}\|_F) := L_c,\end{aligned} \tag{12.15}$$

where the inequality follows from $|\langle \mathbf{A},\mathbf{B}\rangle| \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_F$ and $\|\mathbf{CD}\|_F \leq \|\mathbf{C}\|_F\|\mathbf{D}\|_F$. Due to the constraints on $\mathbf{U}$ and $\mathbf{V}$, we have $\|\mathbf{U}\|_F^2 = \mathrm{tr}(\mathbf{U}^\top\mathbf{U}) = \mathrm{tr}(\mathbf{I}) = r$, $\|\mathbf{V}\|_F^2 = \sum_{j=1}^n \|\mathbf{v}_j\|^2 = n$. □

To analyse the convergence, we rewrite Eq. (12.6) as

$$\underset{\mathbf{U},\mathbf{V}}{\text{minimize}}\, f(\mathbf{U},\mathbf{V}) = h(\mathbf{U},\mathbf{V}) + \delta_\mathbb{U}(\mathbf{U}) + \delta_\mathbb{V}(\mathbf{V}), \tag{12.16}$$

where

$$\delta_\mathbb{U}(\mathbf{U}) = \begin{cases} 0, & \mathbf{U}\in\mathbb{U} \\ \infty, & \mathbf{U}\notin\mathbb{U}\end{cases}$$

is the indicator function of the set $\mathbb{U}$ and therefore nonsmooth, so is $\delta_\mathbb{V}(\mathbf{V})$.

The following result establishes that the subsequence convergence property of the proposed algorithm, i.e., the sequence generated by Algorithm 9 is bounded and any of its limit point is a critical point of Eq. (12.16).

Theorem 12.4.1 (Subsequence convergence)**.** *Let $\{\mathbf{W}(k)\}_{k\geq 0} = \{(\mathbf{U}(k),\mathbf{V}(k))\}_{k\geq 0}$ be the sequence generated by Algorithm 9 with constant step size $\lambda,\mu > L_c$. Then the sequence $\{\mathbf{W}(k)\}_{k\geq 0}$ is bounded and obeys the following properties:*

(P1): Sufficient decrease:

$$f(\mathbf{W}(k-1)) - f(\mathbf{W}(k)) \geq \frac{\min(\lambda,\mu)-L_c}{2}\|\mathbf{W}(k)-\mathbf{W}(k-1)\|_F^2, \tag{12.17}$$

which implies that

$$\lim_{k\to\infty} \|\mathbf{W}(k-1)-\mathbf{W}(k)\|_F = 0. \tag{12.18}$$

(P2): The sequence $\{f(\mathbf{W}(k))\}_{k\geq 0}$ is convergent.

(P3): For any convergent subsequence $\{\mathbf{W}(k')\}$, its limit point $\mathbf{W}^\star$ is a critical point of f and

$$\lim_{k'\to\infty} f(\mathbf{W}(k')) = \lim_{k\to\infty} f(\mathbf{W}(k)) = f(\mathbf{W}^\star). \tag{12.19}$$

Proof of Theorem 12.4.1. Before proving Theorem 12.4.1, we give out some necessary definition.

Definition 12.4.1. *[257] Let $f:\mathbb{R}^d \to (-\infty,\infty]$ be a proper and lower semi-continuous function, whose domain is defined as*

$$\operatorname{dom} f := \{\mathbf{u} \in \mathbb{R}^n : f(\mathbf{u}) < \infty\}.$$

The (Fréchet) subdifferential ∂f of f at $\mathbf{u}$ is defined by

$$\partial f(\mathbf{u}) = \left\{ \mathbf{z} \in \mathbb{R}^d : \lim_{\mathbf{v} \to \mathbf{u}} \inf \frac{f(\mathbf{v}) - f(\mathbf{u}) - \langle \mathbf{z}, \mathbf{v} - \mathbf{u} \rangle}{\|\mathbf{u} - \mathbf{v}\|} \geq 0 \right\}$$

for any $\mathbf{u} \in \operatorname{dom} h$ and $\partial f(\mathbf{u}) = \emptyset$ if $\mathbf{u} \notin \operatorname{dom} f$.

We say $\mathbf{u}$ is a limiting critical point, or simply a critical point of f if

$$\mathbf{0} \in \partial f(\mathbf{u}).$$

We now turn to prove Theorem 12.4.1.

- **Showing (P1):** First note that for all k, according to our alternating minimization method, we always have $\delta_{\mathbb{U}}(\mathbf{U}(k)) = \delta_{\mathbb{V}}(\mathbf{V}(k)) = 0$ and thus $f(\mathbf{W}(k)) = h(\mathbf{W}(k))$.

  Since $h(\mathbf{U}, \mathbf{V})$ has Lipschitz continuous gradient at $\mathbf{U} \in \mathbb{U}, \mathbf{V} \in \mathbb{V}$ with Lipschitz gradient L_c and $\lambda > L_c$, we define $h_{L_c}(\mathbf{U}, \mathbf{U}', \mathbf{V})$ as proximal regularization of $h(\mathbf{U}, \mathbf{V})$ linearized at $\mathbf{U}', \mathbf{V}$:

$$h(\mathbf{U}', \mathbf{V}) + \langle \nabla_{\mathbf{U}} h(\mathbf{U}', \mathbf{V}), \mathbf{U} - \mathbf{U}' \rangle + \frac{L_c}{2} \|\mathbf{U} - \mathbf{U}'\|_F^2,$$

  By the definition of Lipschitz continuous gradient and Taylor expansion, we have

$$h(\mathbf{U}, \mathbf{V}) \leq h_{L_c}(\mathbf{U}, \mathbf{U}', \mathbf{V}). \tag{12.20}$$

  Also by the definition of proximal map, we get:

$$\mathbf{U}(k) = \arg\min_{\mathbf{U}} \delta_{\mathbb{U}}(\mathbf{U}) + \frac{\mu}{2} \|\mathbf{U} - \mathbf{U}(k-1)\|_F^2 + \langle \nabla_{\mathbf{U}} h(\mathbf{U}(k-1), \mathbf{V}(k-1)), \mathbf{U} - \mathbf{U}(k-1) \rangle \tag{12.21}$$

  and hence we take $\mathbf{U}(k) = \mathbf{U}$, which implies that

$$\delta_{\mathbb{U}}(\mathbf{U}(k)) + \frac{\mu}{2} \|\mathbf{U}(k) - \mathbf{U}(k-1)\|_F^2 + \langle \nabla_{\mathbf{U}} h(\mathbf{U}(k-1), \mathbf{V}(k-1)), \mathbf{U}(k) - \mathbf{U}(k-1) \rangle \leq \delta_{\mathbb{U}}(\mathbf{U}(k-1)) \tag{12.22}$$

Combining Eq. (12.20) and Eq. (12.22), we have:

$$\begin{aligned}
&h(\mathbf{U}(k),\mathbf{V}(k-1))+\delta_{\mathbb{U}}(\mathbf{U}(k))\\
&\leq h(\mathbf{U}(k-1),\mathbf{V}(k-1))+\langle\nabla_{\mathbf{U}}h(\mathbf{U}(k-1),\mathbf{V}(k-1)),\mathbf{U}(k)-\mathbf{U}(k-1)\rangle+\frac{L_c}{2}\|\mathbf{U}(k)-\mathbf{U}(k-1)\|_F^2\\
&\quad+\delta_{\mathbb{U}}(\mathbf{U}(k))\\
&\leq h(\mathbf{U}(k-1),\mathbf{V}(k-1))+\frac{L_c}{2}\|\mathbf{U}(k)-\mathbf{U}(k-1)\|_F^2+\delta_{\mathbb{U}}(\mathbf{U}(k-1))-\frac{\mu}{2}\|\mathbf{U}(k)-\mathbf{U}(k-1)\|_F^2\\
&=h(\mathbf{U}(k-1),\mathbf{V}(k-1))+\delta_{\mathbb{U}}(\mathbf{U}(k-1))-\frac{\mu-L_c}{2}\|\mathbf{U}(k)-\mathbf{U}(k-1)\|_F^2,
\end{aligned}\tag{12.23}$$

Similarly, we have

$$h(\mathbf{U}(k),\mathbf{V}(k))-h(\mathbf{U}(k),\mathbf{V}(k-1))+\delta_{\mathbb{V}}(\mathbf{V}(k))-\delta_{\mathbb{V}}(\mathbf{V}(k-1))\leq-\frac{\lambda-L_c}{2}\|\mathbf{V}(k)-\mathbf{V}(k-1)\|_F^2\tag{12.24}$$

which together with the above equation gives Eq. (12.17). Now repeating Eq. (12.17) for all k will give

$$(\min(\lambda,\mu)-L_c)\sum_{k=1}^{\infty}\|\mathbf{W}(k)-\mathbf{W}(k-1)\|_F^2\leq f(\mathbf{W}(0)),\tag{12.25}$$

which gives Eq. (12.18).

Remark 12.4.1. In our proposed algorithm, since in every update, our solution is closed while satisfying the constraints, thus in fact $\delta_{\mathbb{U}}$ and $\delta_{\mathbb{V}}$ are 0, and ∞ is never achieved.

- **Showing (P2):** It follows from Eq. (16) that $\{f(\mathbf{W}(k))\}_{k\geq 0}$ is a decreasing sequence. Due to the fact that f is lower bounded as $f(\mathbf{W}(k))\geq 0$ for all k, we conclude that $\{f(\mathbf{W}(k))\}_{k\geq 0}$ is convergent.

- **Showing (P3):** Since $\mathbf{U}(k')\in\mathbb{U},\mathbf{V}(k')\in\mathbb{V}$ for all k' and both of the sets $\mathbb{U}$ and $\mathbb{V}$ are closed, we have $\mathbf{U}^\star\in\mathbb{U},\mathbf{V}^\star\in\mathbb{V}$. Since h is continuous, we have

$$\lim_{k'\to\infty}f(\mathbf{W}(k'))=\lim_{k'\to\infty}h(\mathbf{U}(k'),\mathbf{V}(k'))+\delta_{\mathbb{U}}(\mathbf{U}(k'))+\delta_{\mathbb{V}}(\mathbf{V}(k'))=f(\mathbf{W}^\star)$$

which together with the fact that $\{f(\mathbf{W}(k))\}_{k\geq 0}$ is convergent gives Eq. (12.18).

To show $\mathbf{W}^\star$ is a critical point, we first consider Eq. (12.21) and the optimality condition yields:

$$\nabla_{\mathbf{U}}h(\mathbf{U}(k-1),\mathbf{V}(k-1))+\mu(\mathbf{U}(k)-\mathbf{U}(k-1))+\partial\delta_{\mathbb{U}}(\mathbf{U}(k))=0.\tag{12.26}$$

Similarly, we have

$$\nabla_{\mathbf{V}}h(\mathbf{U}(k),\mathbf{V}(k-1))+\lambda(\mathbf{V}(k)-\mathbf{V}(k-1))+\partial\delta_{\mathbb{V}}(\mathbf{V}(k))=0.\tag{12.27}$$

Now, define

$$\mathbf{A}_k := \nabla_{\mathbf{U}} h(\mathbf{U}(k), \mathbf{V}_k) + \partial \delta_{\mathbb{U}}(\mathbf{U}(k)),$$
$$\mathbf{B}_k := \nabla_{\mathbf{V}} h(\mathbf{U}(k), \mathbf{V}(k)) + \partial \delta_{\mathbb{V}}(\mathbf{V}(k)).$$

Thus, we have

$$\mathbf{A}_k \in \partial_{\mathbf{U}} f(\mathbf{U}(k), \mathbf{V}(k)), \mathbf{B}_k \in \partial_{\mathbf{V}} f(\mathbf{U}(k), \mathbf{V}(k)). \tag{12.28}$$

It follows from the above that

$$\begin{aligned} \lim_{k\to\infty} \|\mathbf{A}_k\|_F &\leq \lim_{k\to\infty} \|\nabla_{\mathbf{U}} h(\mathbf{U}(k), \mathbf{V}(k)) - \nabla_{\mathbf{U}} h(\mathbf{U}(k-1), \mathbf{V}(k-1))\|_F + \mu \|\mathbf{U}(k) - \mathbf{U}(k-1)\|_F \\ &\leq \lim_{k\to\infty} (L_c + \mu) \|\mathbf{W}(k) - \mathbf{W}(k-1)\|_F = 0. \end{aligned} \tag{12.29}$$

Similarly, we have:

$$\lim_{k\to\infty} \|\mathbf{B}_k\|_F \leq \lim_{k\to\infty} (L_c + \lambda) \|\mathbf{W}(k) - \mathbf{W}(k-1)\|_F = 0. \tag{12.30}$$

Then we have:

$$\operatorname{dist}(\mathbf{0}, \partial f(\mathbf{W}(k))) \leq (2L_c + \mu + \lambda) \|\mathbf{W}(k) - \mathbf{W}(k-1)\|_F \tag{12.31}$$

Owing to the closedness properties of $\partial f(\mathbf{W}(k'))$, we finally obtain

$$\mathbf{0} \in \partial f(\mathbf{W}^\star).$$

Thus, $\mathbf{W}^\star$ is a critical point of f.

□

Theorem 12.4.2 (Sequence convergence)**.** *The sequence $\{\mathbf{W}(k)\}_{k\geq 0}$ generated by Algorithm 9 with a constant step size $\lambda, \mu > L_c$ is global-sequence convergence.*

Proof of Theorem 12.4.2. Before proving Theorem 12.4.2, we give out another important definition.

Definition 12.4.2 (**Kurdyka-Lojasiewicz (KL) property).** *[258] We say a proper semi-continuous function $h(\mathbf{u})$ satisfies Kurdyka-Lojasiewicz (KL) property, if $\overline{\mathbf{u}}$ is a critical point of $h(\mathbf{u})$, then there exist $\delta > 0,\ \theta \in [0, 1),\ C_1 > 0$ such that*

$$|h(\mathbf{u}) - h(\overline{\mathbf{u}})|^\theta \leq C_1 \operatorname{dist}(\mathbf{0}, \partial h(\mathbf{u})),\ \ \forall\, \mathbf{u} \in B(\overline{\mathbf{u}}, \delta)$$

We mention that the above KL property (also known as KL inequality) states the regularity of $h(\mathbf{u})$ around its critical point $\mathbf{u}$ and the KL inequality trivially holds at non-critical point. There are a very large set of functions

satisfying the KL inequality including any semi-algebraic functions [257]. Clearly, the objective function f is semi-algebraic as both h, $\delta_{\mathbb{U}}$ and $\delta_{\mathbb{V}}$ are semi-algebraic.

Lemma 12.4.1 (Uniform KL property)**.** *There exist* $\delta_0 > 0$, $\theta_{KL} \in [0, 1)$, $C_{KL} > 0$ *such that for all* W *s.t.* $\operatorname{dist}(\mathbf{W}, \mathbb{C}(\mathbf{W}(0))) \leq \delta_0$*:*

$$\left|f(\mathbf{W}) - \overline{f}\right|^{\theta_{KL}} \leq C_{KL} \operatorname{dist}(\mathbf{0}, \partial f(\mathbf{W})) \tag{12.32}$$

with $\overline{f}$ *denoting the limiting function value defined in (P2) of Theorem 12.4.1.*

Proof. First we recognize the union $\bigcup_i B(\mathbf{W}_i^\star, \delta_i)$ forms an open cover of $\mathbb{C}(\mathbf{W}(0))$ with $\mathbf{W}_i^\star$ representing all points in $\mathbb{C}(\mathbf{W}(0))$ and δ_i to be chosen so that the the following KL property of f at $\mathbf{W}_i^\star \in \mathbb{C}(\mathbf{W}(0))$ holds:

$$\left|f(\mathbf{W}) - \overline{f}\right|^{\theta_i} \leq C_i \operatorname{dist}(\mathbf{0}, \partial f(\mathbf{W})) \ \ \forall\, \mathbf{W} \in B(\mathbf{W}_i^\star, \delta_i)$$

where we have used all $f(\mathbf{W}_i^\star) = \overline{f}$ by assertion (P3) of Theorem 12.4.1. Then due to the compactness of the set $\mathbb{C}(\mathbf{W}(0))$, it has a finite subcover $\bigcup_{i=1}^p B(\mathbf{W}_{k_i}^\star, \delta_{k_i})$ for some positive integer p. Now combining all, we have for all $W \in \bigcup_{i=1}^p B(\mathbf{W}_{k_i}^\star, \delta_{k_i})$,

$$\left|f(\mathbf{W}) - \overline{f}\right|^{\theta_{KL}} \leq C_{KL} \operatorname{dist}(\mathbf{0}, \partial f(\mathbf{W})) \tag{12.33}$$

with $\theta_{KL} = \max_{i=1}^p \{\theta_{k_i}\}$ and $C_{KL} = \max_{i=1}^p \{C_{k_i}\}$. Finally, since $\bigcup_{i=1}^p B(\mathbf{W}_{k_i}^\star, \delta_{k_i})$ is an open cover of $\mathbb{C}(\mathbf{W}(0))$, there exists a sufficiently small number δ_0 so that

$$\{(\mathbf{W}) : \operatorname{dist}(\mathbf{W}, \mathbb{C}(\mathbf{W}(0))) \leq \delta_0\} \subset \bigcup_{i=1}^p B(\mathbf{W}_i^\star, \delta_{k_i}).$$

Therefore, eq. (12.33) holds whenever $\operatorname{dist}(\mathbf{W}, \mathbb{C}(\mathbf{W}(0))) \leq \delta_0$. □

We now turn to prove Theorem 12.4.2.

According to Definition 12.4.2, there exists a sufficiently large k_0 satisfying:

$$[f(\mathbf{W}(k)) - f(\mathbf{W}^\star)]^\theta \leq C_2 \operatorname{dist}(\mathbf{0}, \partial f(\mathbf{W}(k))), \ \ \forall k \geq k_0. \tag{12.34}$$

In the subsequent analysis, we restrict to $k \geq k_0$. Construct a concave function $x^{1-\theta}$ for some $\theta \in [0, 1)$ with domain $x > 0$. Obviously, by the concavity, we have

$$x_2^{1-\theta} - x_1^{1-\theta} \geq (1-\theta) x_2^{-\theta} (x_2 - x_1), \forall x_1 > 0, x_2 > 0$$

Replacing x_1 by $f(\mathbf{W}_{k+1}) - f(\mathbf{W}^\star)$ and x_2 by $f(\mathbf{W}_k) - f(\mathbf{W}^\star)$ and using the sufficient decrease property, we have

$$
\begin{aligned}
&[f(\mathbf{W}(k)) - f(\mathbf{W}^\star)]^{1-\theta} - [f(\mathbf{W}(k+1)) - f(\mathbf{W}^\star)]^{1-\theta} \\
&\geq (1-\theta)\frac{f(\mathbf{W}(k)) - f(\mathbf{W}(k+1))}{[f(\mathbf{W}(k)) - f(\mathbf{W}^\star)]^{\theta}} \\
&\geq \frac{\lambda(1-\theta)}{2C_2}\frac{\|\mathbf{W}(k) - \mathbf{W}(k+1)\|_F^2}{\operatorname{dist}(\mathbf{0}, \partial f(\mathbf{W}(k)))}, \\
&\geq \frac{\lambda(1-\theta)}{2C_2C_3}\frac{\|\mathbf{W}(k) - \mathbf{W}(k+1)\|_F^2}{\|\mathbf{W}(k) - \mathbf{W}(k-1)\|_F} \\
&= \kappa(\frac{\|\mathbf{W}(k) - \mathbf{W}(k+1)\|_F^2}{\|\mathbf{W}(k) - \mathbf{W}(k-1)\|_F} + \|\mathbf{W}(k) - \mathbf{W}(k-1)\|_F) - \kappa\|\mathbf{W}(k) - \mathbf{W}(k-1)\|_F \\
&\geq \kappa\left(2\|\mathbf{W}(k) - \mathbf{W}(k+1)\|_F - \|\mathbf{W}(k) - \mathbf{W}(k-1)\|_F\right)
\end{aligned}
$$

And accordingly, we have:

$$
2\|\mathbf{W}(k) - \mathbf{W}(k+1)\|_F - \|\mathbf{W}(k) - \mathbf{W}(k-1)\|_F \leq \beta\left([f(\mathbf{W}(k)) - f(\mathbf{W}^\star)]^{1-\theta} - [f(\mathbf{W}(k+1)) - f(\mathbf{W}^\star)]^{1-\theta}\right) \tag{12.35}
$$

with $C_3 := 2L_c + \mu + \lambda, \kappa := \frac{\lambda(1-\theta)}{2C_2C_3}, \beta := \left(\frac{\lambda(1-\theta)}{2C_2C_3}\right)^{-1}$.

Summing the above inequalities up from some $\widetilde{k} > k_0$ to infinity yields

$$
\sum_{k=\widetilde{k}}^{\infty} \|\mathbf{W}(k) - \mathbf{W}(k+1)\|_F \leq \|\mathbf{W}(\widetilde{k}) - \mathbf{W}(\widetilde{k}-1)\|_F + \beta[f(\mathbf{W}(\widetilde{k})) - f(\mathbf{W}^\star)]^{1-\theta} \tag{12.36}
$$

implying

$$
\sum_{k=\widetilde{k}}^{\infty} \|\mathbf{W}(k) - \mathbf{W}(k+1)\|_F < \infty.
$$

Following some standard arguments one can see that

$$
\limsup_{t\to\infty, t_1, t_2 \geq t} \|\mathbf{W}(t_1) - \mathbf{W}(t_2)\|_F = 0
$$

which implies that the sequence $\{\mathbf{W}(k)\}$ is Cauchy, and hence convergent. Hence, the limit point set $\mathcal{C}(\mathbf{W}(0))$ is singleton $\mathbf{W}^\star$. □

Theorem 12.4.3 (Convergence Rate)**.** *The convergence rate is at least sub-linear.*

Towards that end, we first know from the above argument that $\{\mathbf{W}(k)\}$ converges to some point $\mathbf{W}^\star$, i.e., $\lim_{k\to\infty} \mathbf{W}(k) = \mathbf{W}^\star$. Then using Equation (12.36) and the triangle inequality, we obtain

$$
\begin{aligned}
\|\mathbf{W}(\widetilde{k}) - \mathbf{W}^\star\|_F &\leq \sum_{k=\widetilde{k}}^{\infty} \|\mathbf{W}(k) - \mathbf{W}(k+1)\|_F \\
&\leq \|\mathbf{W}(\widetilde{k}) - \mathbf{W}(\widetilde{k}-1)\|_F + \beta[f(\mathbf{W}(\widetilde{k})) - f(\mathbf{W}^\star)]^{1-\theta}
\end{aligned} \tag{12.37}
$$

which indicates the convergence rate of $\mathbf{W}(\widetilde{k}) \to \mathbf{W}^\star$ is at least as fast as the rate that $\|\mathbf{W}(\widetilde{k}) - \mathbf{W}(\widetilde{k}-1)\|_F + \beta[f(\mathbf{W}(\widetilde{k})) - f(\mathbf{W}^\star)]^{1-\theta}$ converges to 0. In particular, the second term $\beta[f(\mathbf{W}(\widetilde{k})) - f(\mathbf{W}^\star)]^{1-\theta}$ can be controlled:

$$\begin{aligned}\beta[f(\mathbf{W}(\widetilde{k})) - f(\mathbf{W}^\star)]^\theta &\leq \beta C_2 \operatorname{dist}(\mathbf{0}, \partial f(\mathbf{W}(\widetilde{k}))) \\ &\leq \underbrace{\beta C_2(2B_0 + \lambda + \|\mathbf{X}\|_F)}_{:=\alpha} \|\mathbf{W}(\widetilde{k}) - \mathbf{W}(\widetilde{k}-1)\|_F\end{aligned} \tag{12.38}$$

Plugging (12.38) back to (12.37), we then have

$$\sum_{k=\widetilde{k}}^{\infty} \|\mathbf{W}(k) - \mathbf{W}(k+1)\|_F \leq \|\mathbf{W}(\widetilde{k}) - \mathbf{W}(\widetilde{k}-1)\|_F + \alpha \|\mathbf{W}(\widetilde{k}) - \mathbf{W}(\widetilde{k}-1)\|_F^{\frac{1-\theta}{\theta}}.$$

We divide the following analysis into two cases based on the value of the KL exponent θ.

- *Case I*: If $\theta = 0$, we set $\mathcal{Q} := \{k \in \mathbb{N} : \mathbf{W}(k+1) \neq \mathbf{W}(k)\}$ and take k in $\mathcal{Q}$. When k is sufficiently large, then we have:

$$\|\mathbf{W}(k+1) - \mathbf{W}(k)\|_F^2 := C_4 > 0 \tag{12.39}$$

  On the other hand,

$$\begin{aligned}f(\mathbf{W}(k+1)) - f(\mathbf{W}(k)) &\geq \frac{\min(\lambda, \mu) - L_c}{2} \|\mathbf{W}(k+1) - \mathbf{W}(k)\|_F^2 \\ &= \frac{\min(\lambda, \mu) - L_c}{2} C_4\end{aligned} \tag{12.40}$$

  Since $f(\mathbf{W}(k))$ is known to be converged to 0, Eq. (12.40) implies that $\mathcal{Q}$ is finite and sequence $\mathbf{W}(k)$ converges in a finite number of steps.

- *Case II*: $\theta \in (0, \frac{1}{2}]$. This case means $\frac{1-\theta}{\theta} \geq 1$. We define $P_{\widetilde{k}} = \sum_{i=\widetilde{k}}^{\infty} \|\mathbf{W}_{i+1} - \mathbf{W}_i\|_F$,

$$P_{\widetilde{k}} \leq P_{\widetilde{k}-1} - P_{\widetilde{k}} + \alpha \left[P_{\widetilde{k}-1} - P_{\widetilde{k}}\right]^{\frac{1-\theta}{\theta}}. \tag{12.41}$$

  Since $P_{\widetilde{k}-1} - P_{\widetilde{k}} \to 0$, there exists a positive integer k_1 such that $P_{\widetilde{k}-1} - P_{\widetilde{k}} < 1,\ \forall\, \widetilde{k} \geq k_1$. Thus,

$$P_{\widetilde{k}} \leq (1+\alpha)\,(P_{\widetilde{k}-1} - P_{\widetilde{k}}), \quad \forall\, \widetilde{k} \geq \max\{k_0, k_1\},$$

  which implies that

$$P_{\widetilde{k}} \leq \rho \cdot P_{\widetilde{k}-1}, \quad \forall\, \widetilde{k} \geq \max\{k_0, k_1\}, \tag{12.42}$$

  where $\rho = \frac{1+\alpha}{2+\alpha} \in (0,1)$. This together with (12.37) gives the linear convergence rate

$$\|\mathbf{W}(k) - \mathbf{W}^\star\|_F \leq \mathcal{O}(\rho^{k-\overline{k}}),\ \forall\, k \geq \overline{k}. \tag{12.43}$$

  where $\overline{k} = \max\{k_0, k_1\}$.

- *Case III*: $\theta \in (1/2, 1)$. This case means $\frac{1-\theta}{\theta} \leq 1$. Based on the former results, we have

$$P_{\widetilde{k}} \leq (1+\alpha)\left[P_{\widetilde{k}-1} - P_{\widetilde{k}}\right]^{\frac{1-\theta}{\theta}}, \quad \forall\, \widetilde{k} \geq \max\{k_0, k_1\}.$$

We now run into the same situation as in [259]. Hence following a similar argument gives

$$P_{\widetilde{k}}^{\frac{1-2\theta}{1-\theta}} - P_{\widetilde{k}-1}^{\frac{1-2\theta}{1-\theta}} \geq \zeta,\ \forall\, k \geq \overline{k}$$

for some $\zeta > 0$. Then repeating and summing up the above inequality from $\overline{k} = \max\{k_0, k_1\}$ to any $k > \overline{k}$, we can conclude

$$P_{\widetilde{k}} \leq \left[P_{\widetilde{k}-1}^{\frac{1-2\theta}{1-\theta}} + \zeta(\widetilde{k} - \overline{k})\right]^{-\frac{1-\theta}{2\theta-1}} = \mathcal{O}\left((\widetilde{k} - \overline{k})^{-\frac{1-\theta}{2\theta-1}}\right).$$

Finally, the following sublinear convergence holds

$$\|\mathbf{W}(k) - \mathbf{W}^\star\|_F \leq \mathcal{O}\left((k - \overline{k})^{-\frac{1-\theta}{2\theta-1}}\right),\ \forall\, k > \overline{k}. \tag{12.44}$$

We end this proof by commenting that both linear and sublinear convergence rate are closely related to the KL exponent θ at the critical point $\mathbf{W}^\star$.

12.5 Experiments

In this section, we are going to apply our proposed spherical PCA to both synthetic data and real-world datasets to test the performance of our proposed method. The experiment on synthetic data will be introduced first followed by experiments on real-world datasets.

12.5.1 Synthetic Data Experiment

We first generate 200 data points, half of which is distributed within the region between $X = Z$ and Z axis (denoted as blue dots in the top part of Figure 12.2), while another group is generated within the region between $Y = Z$ and Z axis (denoted as the red dots). These two clusters of data are generated through different angles. Thus when we do clustering, it should be angle distance rather than Euclidean distance to determine the clustering result. For our method, we learn a projection matrix $\mathbf{U} \in \mathbb{R}^{3\times 2}$ and plot the component matrix $\mathbf{V} \in \mathbb{R}^{2\times 200}$ as the bottom part illustrates. We see that, Euclidean distance-based method (such as K-means) will yield poor clustering result (middle part), while spherical-PCA will obtain good clustering result.

Also, we show the convergence of $\{\mathbf{W}(k)\}_{k\geq 0} = \{(\mathbf{U}(k), \mathbf{V}(k))\}_{k\geq 0}$ generated by our method. As Figure 12.3 shows, after short iterations, the generated sequences will be stable, which is in accordance with the convergence proof. It also illustrates the objective with update. We see that it converges fast with a sublinear rate, which validates our convergence rate analysis.

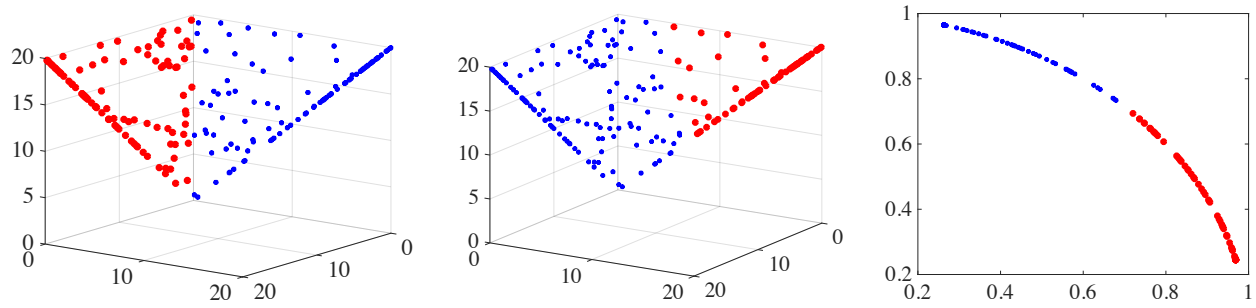


Figure 12.2: **Left**: two groups of data generated from two angles. **Middle**: clustering result with distance -based method K-means. **Right**: clustering result with our method. Blue and red represent different clusters.

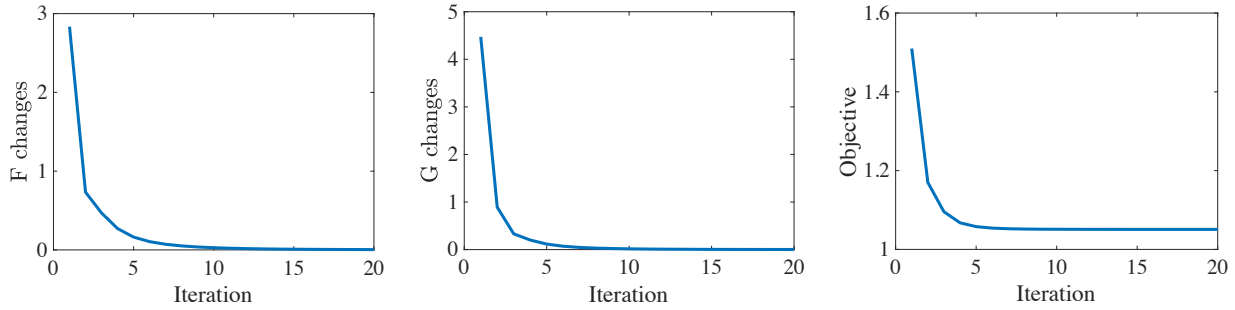


Figure 12.3: **Left**: $\|\mathbf{U}(k+1)-\mathbf{U}(k)\|_F$ with updates. **Center**: $\|\mathbf{V}(k+1)-\mathbf{V}(k)\|_F$ with updates. Both converge to 0 after several iterations. **Right**: Objective converges at sub-linear rate. All validate our analysis.

Table 12.1: Clustering performance of different algorithms on 20-newsgroup dataset

Methods	K-means		MUA		PCA		R1-PCA		K-SVD		**Spherical PCA**	
#Groups	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI
5	0.651	0.621	0.674	0.614	0.703	0.628	0.745	0.647	0.789	0.673	**0.838**	**0.695**
10	0.487	0.316	0.478	0.320	0.502	0.383	0.535	0.398	0.527	0.394	**0.588**	**0.401**
15	0.398	0.307	0.387	0.301	0.412	0.319	0.423	0.320	0.461	0.377	**0.486**	**0.385**
20	0.315	0.242	0.314	0.221	0.362	0.248	0.394	0.260	0.412	0.280	**0.431**	**0.294**

Table 12.2: Clustering performance of different algorithms on four UCI datasets

Methods	K-means		MUA		PCA		R1-PCA		K-SVD		**Spherical PCA**	
Data (#class)	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI
glass (6)	0.687	0.566	0.692	0.574	0.732	0.608	0.769	0.626	**0.801**	**0.648**	0.788	0.635
diabetes (2)	0.775	0.632	0.788	0.654	0.761	0.613	0.808	0.631	0.827	0.672	**0.832**	**0.680**
mfeat (10)	0.365	0.223	0.358	0.211	0.371	0.225	**0.431**	**0.342**	0.412	0.328	0.425	0.330
isolet (26)	0.267	0.198	0.253	0.181	0.262	0.182	0.324	0.201	0.357	0.246	**0.373**	**0.250**

12.5.2 Real-world Datasets Experiment

It is known that in information retrieval, similarities or dissimilarities (proximities) between objects are more critical than Euclidean distance. In this subsection, we will test our proposed method on the widely-used 20-newsgroup dataset for clustering. We have different newsgroups such as: *comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space, talk.politics.mideast*, etc.. 200 documents are randomly sampled from each newsgroup. The word-document matrix X is constructed with 500 words selected according to the mutual information between words and documents. *Tf.idf* term weighting is used before normalization. Clustering accuracy are computed using the known class labels. Results will be compared including clustering accuracy (Acc.) and Normalized Mutual Information (NMI) [260].

Different clustering algorithms will be compared including:

1. **R1-PCA**, which proposes a rotational invariant ℓ_1-norm PCA, where a robust covariance matrix will soften the effects of outliers [261];
2. **K-SVD**, which is an iterative method that alternates between sparse coding of the examples based on the current dictionary and a process of updating the dictionary atoms to better fit the data [262];
3. **PCA**, i.e. the vanilla PCA method in Eq. (12.2) without the constraint on $\mathbf{V}$, which will be Euclidean distance-based by default;
4. **NMF** Matrix Factorization proposed by [263–266] where $\mathbf{U}$ and $\mathbf{V}$ are obtained by Multiplicative Updating Algorithm with nonnegative constraint
5. K**-means** [267].

We vary the number of clusters from 5 to $10, 15$ and 20. In each newsgroup, 200 documents are randomly sampled, and we repeat for 10 times by taking the average and report the clustering result as Table 12.1 demonstrates.

We see that our proposed method Spherical PCA can always achieve both higher clustering accuracy and normalized mutual information in text analysis.

We also compare our method with other methods on UCI datasets including: *glass, diabetes, mfeat* and *isolet*. Table 12.2 illustrates the results. We see that though our method doesn't show the absolute advantage as on text, still the result is considerably good.

All the experiments indicate that our method can achieve good performance on both text and non-text datasets, showing its potential for broader application.

12.6 Conclusion

In this work, we study spherical PCA where the direction matrix is orthonormal and the component vectors are assumed to lie in the unitary sphere. The benefit is obvious that it can make the angle distance equivalent to Euclidean

distance. Due to the nonconvexity of objective function and constraints on the factors which are difficult to tackle, we propose an alternating linearized minimization method to derive the solution, which is proved to be sequence convergent. Moreover, we analyze the convergence rate which is validated by our experiments. The results on real-world datasets and synthetic data illustrate the superiority of our method.

REFERENCES CITED

[1] Qiuwei Li and Gongguo Tang. Approximate support recovery of atomic line spectral estimation: A tale of resolution and precision. In *Signal and Information Processing (GlobalSIP), 2016 IEEE Global Conference on*, pages 153–156. IEEE, 2016.

[2] Qiuwei Li and Gongguo Tang. Approximate support recovery of atomic line spectral estimation: A tale of resolution and precision. *Applied and Computational Harmonic Analysis*, 2018.

[3] Qiuwei Li, Ashley Prater, Lixin Shen, and Gongguo Tang. Overcomplete tensor decomposition via convex optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 53–56. IEEE, 2015.

[4] Qiuwei Li, Ashley Prater, Lixin Shen, and Gongguo Tang. A super-resolution framework for tensor decomposition. *arXiv preprint arXiv:1602.08614*, 2016.

[5] Qiuwei Li and Gongguo Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1235–1239. IEEE, 2017.

[6] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 03 2018.

[7] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. Global optimality in low-rank matrix optimization. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1275–1279, Nov 2017.

[8] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

[9] Qiuwei Li, Zhihui Zhu, Gongguo Tang, and Michael B Wakin. The geometry of equality-constrained global consensus problems. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7928–7932. IEEE, 2019.

[10] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. Alternating minimizations converge to second-order optimal solutions. In *International Conference on Machine Learning*, pages 3935–3943, 2019.

[11] Qiuwei Li, Zhihui Zhu, Gongguo Tang, and Michael B Wakin. Provable bregman-divergence based methods for nonconvex and non-lipschitz problems. *arXiv preprint arXiv:1904.09712*, 2019.

[12] Kai Liu, Qiuwei Li, Hua Wang, and Gongguo Tang. Spherical principal component analysis. *arXiv preprint arXiv:1903.06877*, 2019.

[13] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, June 2014.

[14] Emmanuel J Candès and Carlos Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.

[15] Carlos Fernandez-Granda. Super-resolution of point sources via convex programming. *Information and Inference: A Journal of the IMA*, 5(3):251–303, 2016.

[16] Gongguo Tang, B N Bhaskar, P Shah, and B Recht. Compressed Sensing off the Grid. *Information Theory, IEEE Transactions on*, 59(11):7465–7490, 2013.

[17] Gongguo Tang, Badri Narayan Bhaskar, and Benjamin Recht. Near minimax line spectral estimation. *IEEE Transactions on Information Theory*, 61(1):499–512, 2015.

[18] G. Tang, P. Shah, B. N. Bhaskar, and B. Recht. Robust line spectral estimation. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 301–305, Nov 2014.

[19] Carlos Fernandez-Granda, Gongguo Tang, Xiaodong Wang, and Le Zheng. Demixing sines and spikes: Robust spectral super-resolution in the presence of outliers. *Information and Inference: A Journal of the IMA*, page iax005, 2016.

[20] Petre Stoica and Nehorai Arye. MUSIC, maximum likelihood, and Cramer-Rao bound. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(5):720–741, May 1989.

[21] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[22] Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999, 2013.

[23] Nikhil Rao, Parikshit Shah, and Stephen Wright. Forward–backward greedy algorithms for atomic norm regularization. *IEEE Transactions on Signal Processing*, 63(21):5798–5811, 2015.

[24] Ambuj Tewari, Pradeep K Ravikumar, and Inderjit S Dhillon. Greedy algorithms for structurally constrained high dimensional problems. In *Advances in Neural Information Processing Systems*, pages 882–890, 2011.

[25] Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*, pages 57–60. IEEE, 2015.

[26] Armin Eftekhari and Michael B Wakin. Greed is super: A new iterative method for super-resolution. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 631–631. IEEE, 2013.

[27] Gongguo Tang. Resolution limits for atomic decompositions via Markov-Bernstein type inequalities. In *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pages 548–552, Washington, DC, May 2015. IEEE.

[28] Carlos Fernandez-Granda. Support detection in super-resolution. In *Sampling Theory and Applications (SampTA), 2013 International Conference on*, pages 145–148. IEEE, 2013.

[29] Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

[30] Quentin Denoyelle, Vincent Duval, and Gabriel Peyré. Support recovery for sparse super-resolution of positive measures. *Journal of Fourier Analysis and Applications*, 23(5):1153–1194, Oct 2017.

[31] Veniamin I Morgenshtern and Emmanuel J Candes. Super-resolution of positive sources: The discrete setup. *SIAM Journal on Imaging Sciences*, 9(1):412–444, 2016.

[32] Steven M Kay. *Modern Spectral Estimation*. Pearson Education India, 1988.

[33] Petre Stoica and Randolph L Moses. *Introduction to Spectral Analysis*, volume 1. Prentice Hall Upper Saddle River, 1997.

[34] R de Prony. Essai experimental et analytique. *J. Ec. Polytech.(Paris)*, 2:24–76, 1795.

[35] MH Kahn, MS Mackisack, MR Osborne, and GK Smyth. On the consistency of prony's method and related algorithms. *Journal of Computational and Graphical Statistics*, 1(4):329–349, 1992.

[36] Y Hua and T K Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 38(5):814–824, May 1990.

[37] R Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. on Antennas and Propagation*, 34(3):276–280, 1986.

[38] R Roy and T Kailath. ESPRIT - estimation of signal parameters via rotational invariance techniques. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37(7):984–995, 1989.

[39] Martin Vetterli, Pina Marziliano, and Thierry Blu. Sampling signals with finite rate of innovation. *IEEE Trans. on Signal Processing*, 50(6):1417–1428, 2002.

[40] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[41] Emmanuel J Candès et al. Compressive sampling. In *Proceedings of the international congress of mathematicians*, volume 3, pages 1433–1452. Madrid, Spain, 2006.

[42] Richard G Baraniuk. Compressive sensing. *IEEE signal processing magazine*, 24(4), 2007.

[43] Yuejie Chi, Louis L Scharf, Ali Pezeshki, and A Robert Calderbank. Sensitivity to basis mismatch in compressed sensing. *IEEE Transactions on Signal Processing*, 59(5):2182–2195, 2011.

[44] Gongguo Tang, Badri Narayan Bhaskar, and Benjamin Recht. Sparse recovery over continuous dictionaries-just discretize. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047. IEEE, 2013.

[45] Vincent Duval and Gabriel Peyré. Sparse spikes deconvolution on thin grids. *arXiv preprint arXiv:1503.08577*, 2015.

[46] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[47] Yuanxin Li and Yuejie Chi. off-the-grid line spectrum denoising and estimation with multiple measurement vectors. *IEEE Transactions on Signal Processing*, 64(5):1257–1269, 2016.

[48] Zai Yang and Lihua Xie. Exact joint sparse frequency recovery via optimization methods. *IEEE Transactions on Signal Processing*, 64(19):5145–5157, 2014.

[49] Shuang Li, Dehui Yang, Gongguo Tang, and Michael B Wakin. Atomic norm minimization for modal analysis from random and compressed samples. *IEEE Transactions on Signal Processing*, 66(7):1817–1831, 2018.

[50] Jean-Marc Azais, Yohann De Castro, and Fabrice Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.

[51] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (LASSO). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[52] Johann A Bengua, Ho N Phien, Hoang Duong Tuan, and Minh N Do. Efficient tensor completion for color image and video recovery: Low-rank tensor train. *IEEE Transactions on Image Processing*, 26(5):2466–2479, 2017.

[53] Jieqiong Hou and Haifeng Qian. Collaboratively filtering malware infections: A tensor decomposition approach. In *Proceedings of the ACM Turing 50th Celebration Conference-China*, page 28. ACM, 2017.

[54] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

[55] Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018.

[56] Nadav Cohen, Or Sharir, Yoav Levine, Ronen Tamari, David Yakira, and Amnon Shashua. Analysis and design of convolutional networks via hierarchical tensor decompositions. *arXiv preprint arXiv:1705.02302*, 2017.

[57] Age Smilde, Rasmus Bro, and Paul Geladi. *Multi-Way Analysis: Applications in the Chemical Sciences*. John Wiley & Sons, 2005.

[58] Jean-Francois Cardoso. Source separation using higher order moments. *International Conference on Acoustics, Speech, and Signal Processing*, pages 2109–2112 vol.4, 1989.

[59] Navin Goyal, Santosh Vempala, and Ying Xiao. *Fourier PCA and Robust Tensor Decomposition*. ACM, New York, USA, May 2014.

[60] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Analyzing tensor power method dynamics in overcomplete regime. *Journal of Machine Learning Research*, 18(22):1–40, 2017.

[61] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151. ACM, 2015.

[62] Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, pages 1223–1231, 2016.

[63] Joseph B Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95–138, January 1977.

[64] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[65] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, August 2010.

[66] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are NP-Hard. *Journal of the ACM (JACM)*, 60(6):45–39, November 2013.

[67] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[68] Shmuel Friedland and Lek-Heng Lim. Nuclear norm of higher-order tensors. *Mathematics of Computation*, 2017.

[69] Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5249–5257, 2016.

[70] Yaoliang Yu, Hao Cheng, and Xinhua Zhang. Approximate low-rank tensor learning. In *7th NIPS workshop on optimization for machine learning*, volume 1, 2014.

[71] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Learning overcomplete latent variable models through tensor methods. In *Conference on Learning Theory*, pages 36–112, 2015.

[72] Gongguo Tang and Parikshit Shah. Guaranteed tensor decomposition: A moment approach. In *International Conference on Machine Learning*, Lille, France, 2015.

[73] S. Foucart and H. Rauhut. *A Mathematical Introduction To Compressive Sensing*. Applied and Numerical Harmonic Analysis. Springer New York, 2013.

[74] Cristiano Bocci, Luca Chiantini, and Giorgio Ottaviani. Refined methods for the identifiability of tensors. *Annali di Matematica Pura ed Applicata (1923-)*, 193(6):1691–1702, 2014.

[75] Pierre Comon, Xavier Luciani, and André LF De Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 23(7-8):393–405, 2009.

[76] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[77] Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.

[78] Emmanuel J Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, May 2008.

[79] Gongguo Tang and Benjamin Recht. Atomic decomposition of mixtures of translation-invariant signals. In *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing CAMSAP*, Saint Martin, December 2013.

[80] Tamir Bendory, Shai Dekel, and Arie Feuer. Robust recovery of stream of pulses using convex optimization. *Journal of Mathematical Analysis and Applications*, 442(2):511–536, 2016.

[81] Tamir Bendory, Shai Dekel, and Arie Feuer. Super-resolution on the sphere using convex optimization. *Signal Processing, IEEE Transactions on*, 63(9):2253–2262, 2015.

[82] Reinhard Heckel, Veniamin I Morgenshtern, and Mahdi Soltanolkotabi. Super-resolution radar. *Information and Inference: A Journal of the IMA*, 5(1):22–75, 2016.

[83] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[84] Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, February 2011.

[85] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*, pages 73–81, 2014.

[86] Bo Huang, Cun Mu, Donald Goldfarb, and John Wright. Provable low-rank tensor recovery. *Optimization-Online*, 4252, 2014.

[87] Parikshit Shah, Nikhil Rao, and Gongguo Tang. Sparse and low-rank tensor decomposition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2548–2556, 2015.

[88] Alexander Barvinok. *A Course in Convexity*. American Mathematical Soc., 2002.

[89] G A Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, June 1992.

[90] Gongguo Tang, Parikshit Shah, Badri Narayan Bhaskar, and Benjamin Recht. Robust line spectral estimation. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 301–305. IEEE, 2014.

[91] Samuel Burer and Renato D C Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, February 2003.

[92] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. Geometry of factored nuclear norm regularization. *arXiv:1704.01265*, 2017.

[93] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of low-rank matrix optimization. *arXiv preprint arXiv:1703.01256*, 2017.

[94] Jean-Bernard Lasserre. *Moments, Positive Polynomials and Their Applications*. World Scientific, October 2009.

[95] Jiawang Nie. Symmetric tensor nuclear norms. *SIAM Journal on Applied Algebra and Geometry*, 1(1):599–625, 2017.

[96] Stephen Boyd. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[97] Ju Sun. *When Are Nonconvex Optimization Problems Not Scary?* PhD thesis, COLUMBIA UNIVERSITY, 2016.

[98] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, Oct 2018.

[99] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017.

[100] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems 29*, pages 2973–2981, 2016.

[101] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1233–1242, 2017.

[102] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973, 2016.

[103] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *29th Annual Conference on Learning Theory*, pages 530–582, 2016.

[104] Xingguo Li, Zhaoran Wang, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, and Tuo Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*, 2016.

[105] Anastasios Kyrillidis, Amir Kalev, Dohuyng Park, Srinadh Bhojanapalli, Constantine Caramanis, and Sujay Sanghavi. Provable quantum state tomography via non-convex methods. *arXiv preprint arXiv:1711.02524*, 2017.

[106] Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

[107] Dohyung Park, Anastasios Kyrillidis, Srinadh Bhojanapalli, Constantine Caramanis, and Sujay Sanghavi. Provable Burer-Monteiro factorization for a class of norm-constrained matrix problems. *stat*, 1050:1, 2016.

[108] Tuo Zhao, Zhaoran Wang, and Han Liu. Nonconvex low rank matrix factorization via inexact first order oracle. *Advances in Neural Information Processing Systems*, 2015.

[109] Lingxiao Wang, Xiao Zhang, and Quanquan Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. In *Artificial Intelligence and Statistics*, pages 981–990, 2017.

[110] Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *International Conference on Machine Learning*, pages 2332–2341, 2015.

[111] Quoc Tran-Dinh and Zheqi Zhang. Extended Gauss-Newton and Gauss-Newton-ADMM algorithms for low-rank matrix optimization. *arXiv preprint arXiv:1606.03358*, 2016.

[112] Qiuwei Li and Gongguo Tang. Convex and nonconvex geometries of symmetric tensor factorization. In *Signals, Systems and Computers, 2017 Asilomar Conference on*. IEEE, 2017.

[113] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

[114] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — Online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[115] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[116] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

[117] Dennis DeCoste. Collaborative prediction using ensembles of maximum margin matrix factorizations. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 249–256. ACM, 2006.

[118] Pratik Biswas and Yinyu Ye. Semidefinite programming for ad hoc wireless sensor network localization. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 46–54. ACM, 2004.

[119] Samuel Burer and Renato D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, Feb 2003.

[120] Yuanxin Li, Yue Sun, and Yuejie Chi. Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements. *IEEE Transactions on Signal Processing*, 65(2):397–408, 2017.

[121] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[122] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[123] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[124] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[125] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[126] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

[127] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, pages 1–27, 2019.

[128] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077, 2017.

[129] A Saumard and JA Wellner. Log-concavity and strong log-concavity: A review. *Statistics surveys*, 8:45, 2014.

[130] Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 720–727, 2003.

[131] Nicolas Gillis and Francois Glineur. Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1149–1165, 2011.

[132] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

[133] Federica Sciacchitano. *Image reconstruction under non-Gaussian noise*. PhD thesis, Technical University of Denmark (DTU), 2017.

[134] Fadoua Balabdaoui and Jon A Wellner. Chernoff's density is log-concave. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 20(1):231, 2014.

[135] Katta G Murty and Santosh N Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

[136] Eduardo D Sontag and Héctor J Sussmann. Backpropagation can give rise to spurious local minima even for networks without hidden layers. *Complex Systems*, 3(1):91–106, 1989.

[137] Sean R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[138] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

[139] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.

[140] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[141] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

[142] Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.

[143] Ricardo Cabral, Fernando De la Torre, João P Costeira, and Alexandre Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2488–2495, 2013.

[144] Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

[145] Philip Wolfe. Convergence conditions for ascent methods. *SIAM review*, 11(2):226–235, 1969.

[146] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms On Matrix Manifolds*. Princeton University Press, 2009.

[147] Nick Higham and Pythagoras Papadimitriou. Matrix procrustes problems. *Rapport technique, University of Manchester*, 1995.

[148] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science & Business Media, 2 edition, 2006.

[149] Scott Aaronson. The learnability of quantum states. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 463(2088):3089–3114, 2007.

[150] Steven T Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.

[151] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1329–1336, 2004.

[152] Gongguo Tang and Arye Nehorai. Lower bounds on the mean-squared error of low-rank matrix reconstruction. *IEEE Transactions on Signal Processing*, 59(10):4559–4571, 2011.

[153] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[154] Guangcan Liu and Ping Li. Low-rank matrix completion in the presence of high coherence. *IEEE Transactions on Signal Processing*, 64(21):5623–5633, 2016.

[155] Maryam Fazel, Haitham Hindi, and S Boyd. Rank minimization and applications in system theory. In *American Control Conference*, volume 4, pages 3273–3278. IEEE, 2004.

[156] Emmanuel J Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[157] Zaid Harchaoui, Matthijs Douze, Mattis Paulin, Miroslav Dudik, and Jérôme Malick. Large-scale image classification with trace-norm regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3386–3393. IEEE, 2012.

[158] Karthik Mohan and Maryam Fazel. Reweighted nuclear norm minimization with application to system identification. In *Proceedings of the 2010 American Control Conference*, pages 2953–2959. IEEE, 2010.

[159] Samuel Burer and Renato D.C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, Jul 2005.

[160] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, Oct 2018.

[161] Madeleine Udell, Corinne Horn, Reza Zadeh, Stephen Boyd, et al. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

[162] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

[163] Thierry Bouwmans, Necdet Serhat Aybat, and El-hadi Zahzah. *Handbook of Robust Low-Rank and Sparse Matrix Decomposition: Applications in Image and Video Processing*. CRC Press, 2016.

[164] Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

[165] Joseph Salmon, Zachary Harmany, Charles-Alban Deledalle, and Rebecca Willett. Poisson noise reduction with non-local PCA. *Journal of Mathematical Imaging and Vision*, 48(2):279–294, 2014.

[166] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, pages 665–674. ACM, 2013.

[167] Emmanuel J Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[168] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

[169] Liangbei Xu and Mark Davenport. Dynamic matrix recovery from incomplete observations under an exact low-rank constraint. In *Advances in Neural Information Processing Systems*, pages 3585–3593, 2016.

[170] Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust Region methods*. SIAM, 2000.

[171] Sonia A Bhaskar and Adel Javanmard. 1-bit matrix completion under exact low-rank constraint. In *49th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, 2015.

[172] Charles R Johnson, Kazuyoshi Okubo, and Robert Reams. Uniqueness of matrix square roots and an application. *Linear Algebra and Its Applications*, 323(1):51–60, 2001.

[173] Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.

[174] Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(1):503–528, 1989.

[175] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[176] Zhang Liu and Lieven Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

[177] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[178] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.

[179] Ioannis Panageas and Georgios Piliouras. Gradient descent converges to minimizers: The case of non-isolated critical points. *CoRR, abs/1605.00405*, 2016.

[180] Emmanuel J Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[181] Yuxin Chen and Emmanuel Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.

[182] Kiryung Lee, Ning Tian, and Justin Romberg. Fast and guaranteed blind multichannel deconvolution under a bilinear system model. *arXiv preprint arXiv:1610.06469*, 2016.

[183] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv preprint arXiv:1606.04933*, 2016.

[184] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *Conference on Learning Theory (COLT)*, pages 123–137, 2014.

[185] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *arXiv preprint arXiv:1511.04777*, 2015.

[186] Huikang Liu, Man-Chung Yue, and Anthony Man-Cho So. On the estimation performance and convergence rate of the generalized power method for phase synchronization. *arXiv preprint arXiv:1603.00211*, 2016.

[187] Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *arXiv preprint arXiv:1605.08370*, 2016.

[188] Gregory S Chirikjian and Alexander B Kyatkin. *Harmonic Analysis for Engineers and Applied Scientists: Updated and Expanded Edition*. Courier Dover Publications, 2016.

[189] Emmanuel J Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[190] Emmanuel J Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[191] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018.

[192] Yudong Chen and Yuejie Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation. *arxiv:1802.08397*, 2018.

[193] P. Jain and P. Kar. *Non-Convex Optimization for Machine Learning*. Foundations and Trends in Machine Learning Series. Now Publishers, 2017.

[194] Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

[195] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

[196] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *arXiv preprint arXiv:1809.09237*, 2018.

[197] Lin Gu, Deze Zeng, Peng Li, and Song Guo. Cost minimization for big data processing in geo-distributed data centers. *IEEE transactions on Emerging topics in Computing*, 2(3):314–323, 2014.

[198] Gesualdo Scutari, Francisco Facchinei, Lorenzo Lampariello, Stefania Sardellitti, and Peiran Song. Parallel and distributed methods for constrained nonconvex optimization-part ii: applications in communications and machine learning. *IEEE Transactions on Signal Processing*, 65(8):1945–1960, 2017.

[199] Mingyi Hong, Jason D Lee, and Meisam Razaviyayn. Gradient primal-dual algorithm converges to second-order stationary solutions for nonconvex distributed optimization. *arXiv preprint arXiv:1802.08941*, 2018.

[200] Annie I-An Chen. *Fast distributed first-order methods*. PhD thesis, Massachusetts Institute of Technology, 2012.

[201] Duvsan Jakovetić, Joao Xavier, and José MF Moura. Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59(5):1131–1146, 2014.

[202] Jinshan Zeng and Wotao Yin. On nonconvex decentralized gradient descent. *IEEE Transactions on Signal Processing*, 66(11):2834–2848, 2018.

[203] Amir Daneshmand, Gesualdo Scutari, and Vyacheslav Kungurtsev. Second-order guarantees of distributed gradient algorithms. *arXiv preprint arXiv:1809.08694*, 2018.

[204] Yuejie Chi. Low-rank matrix completion [lecture notes]. *IEEE Signal Processing Magazine*, 35(5):178–181, Sept 2018.

[205] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

[206] Maher Nouiehed and Meisam Razaviyayn. Learning deep models: Critical points and local openness. In *6th International Conference on Learning Representations – Workshop Track*, 2018.

[207] Clément W. Royer, Michael O'Neill, and Stephen J. Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Mathematical Programming*, Jan 2019.

[208] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[209] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[210] Aryan Mokhtari, Qing Ling, and Alejandro Ribeiro. Network Newton distributed optimization methods. *IEEE Transactions on Signal Processing*, 65(1):146–161, 2017.

[211] Ivano Notarnicola, Ying Sun, Gesualdo Scutari, and Giuseppe Notarstefano. Distributed big-data optimization via block communications. In *IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5. IEEE, 2017.

[212] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, Jan 2009.

[213] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, Aug 2014.

[214] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

[215] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *arXiv preprint arXiv:1803.07726*, 2018.

[216] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

[217] Jingu Kim and Haesun Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 353–362. IEEE, 2008.

[218] Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

[219] Andrzej Cichocki and Anh-Huy Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.

[220] Joseph A O'Sullivan. Alternating minimization algorithms: from Blahut-Arimoto to expectation-maximization. In *Codes, Curves, and Signals*, pages 173–192. Springer, 1998.

[221] Luigi Grippo and Marco Sciandrone. On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. *Operations research letters*, 26(3):127–136, 2000.

[222] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

[223] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[224] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere. *arXiv preprint arXiv:1504.06785*, 2015.

[225] Stanislav P Ponomarev. Submersions and preimages of sets of measure zero. *Siberian Mathematical Journal*, 28(1):153–163, 1987.

[226] Donald Bamber and Jan PH Van Santen. How many parameters can a model have and still be testable? *Journal of Mathematical Psychology*, 29(4):443–473, 1985.

[227] Michael Shub. *Global Stability of Dynamical Systems*. Springer Science & Business Media, 2013.

[228] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 67. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[229] Klaus Fritzsche and Hans Grauert. *From Holomorphic Functions To Complex Manifolds*, volume 213. Springer Science & Business Media, 2012.

[230] Tosio Kato. *Perturbation Theory For Linear Operators*, volume 132. Springer Science & Business Media, 2013.

[231] Yu-Hong Dai and Yaxiang Yuan. A nonlinear conjugate gradient method with a strong global convergence property. *SIAM Journal on optimization*, 10(1):177–182, 1999.

[232] Clément W Royer and Stephen J Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477, 2018.

[233] Kai Zhong, Ian En-Hsu Yen, Inderjit S Dhillon, and Pradeep K Ravikumar. Proximal quasi-Newton for computationally intensive l1-regularized M-estimators. In *Advances in Neural Information Processing Systems 27*, pages 2375–2383, 2014.

[234] Yurii Nesterov and B.T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, Aug 2006.

[235] Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. In *Advances in Neural Information Processing Systems 30*, 2017.

[236] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems 29*, pages 586–594, 2016.

[237] Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[238] Heinz H Bauschke, Patrick L Combettes, et al. *Convex Analysis And Monotone Operator Theory In Hilbert Spaces*, volume 408. Springer, 2011.

[239] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1):91–129, Feb 2013.

[240] Holger Rauhut, Reinhold Schneider, and Zeljka Stojanac. Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications*, 523:220–262, 2017.

[241] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[242] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2114–2121. IEEE, 2009.

[243] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):208–220, 2013.

[244] Qingquan Song, Hancheng Ge, James Caverlee, and Xia Hu. Tensor completion algorithms in big data analytics. *arXiv preprint arXiv:1711.10105*, 2017.

[245] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[246] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[247] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[248] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[249] Hsieh Hou. A fast recursive algorithm for computing the discrete cosine transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(10):1455–1461, 1987.

[250] Hieu V Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*, pages 709–720. Springer, 2010.

[251] Dengdi Sun, Chris HQ Ding, Bin Luo, and Jin Tang. Angular decomposition. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[252] Patrick L Combettes and Teemu Pennanen. Proximal methods for cohypomonotone operators. *SIAM journal on control and optimization*, 43(2):731–742, 2004.

[253] Alexander Kaplan and Rainer Tichatschke. Proximal point methods and nonconvex optimization. *Journal of global Optimization*, 13(4):389–406, 1998.

[254] R Tyrrell Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116, 1976.

[255] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *International conference on machine learning*, pages 352–360, 2013.

[256] Hua Wang, Feiping Nie, Heng Huang, and Fillia Makedon. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[257] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[258] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[259] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

[260] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

[261] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*, pages 281–288. ACM, 2006.

[262] Michal Aharon, Michael Elad, Alfred Bruckstein, et al. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311, 2006.

[263] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[264] Kai Liu and Hua Wang. High-order co-clustering via strictly orthogonal and symmetric l1-norm nonnegative matrix tri-factorization. In *IJCAI*, pages 2454–2460, 2018.

[265] Kai Liu and Hua Wang. Robust multi-relational clustering via l1-norm symmetric nonnegative matrix factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 397–401, 2015.

[266] Hua Wang, Heng Huang, and Chris Ding. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 279–284. ACM, 2011.

[267] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[268] Simon Foucart and Holger Rauhut. *A Mathematical Introduction To Compressive Sensing*. Springer, 2013.

[269] Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific Belmont, 1999.

[270] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[271] Roger A Horn and Charles R Johnson. *Matrix Mnalysis*. Cambridge University Press, 2012.

APPENDIX A

APPENDICES FOR CHAPTER 2

A.1 Jackson Kernel

For any integer $M > 0$, the Jackson kernel, also known as the squared Fejér kernel, is defined by [16, Eq. (IV.2)] or [13, Eq. (2.3) with $M = f_c/2 + 1$]

$$K(f) = \left[\frac{\sin(\pi M f)}{M \sin(\pi f)}\right]^4. \tag{A.1}$$

The Jackson kernel shows up in the construction of dual polynomials that satisfy the Boundedness and Interpolation properties. The choice of the Jackson kernel is due to its nice properties as easily seen from its graph: it attains one at the peak, and quickly decrease to zero. Candès and Fernandez-Granda showed in [13] that as long as the frequencies composing a signal satisfy certain separation condition, then a dual polynomial can be constructed as a linear combination of shifted copies of the Jackson kernel and its first-order derivative to certify that the decomposition achieves the signal's atomic norm.

We use $K'(f), K''(f), K'''(f)$ to denote respectively the first, second, and third order derivatives of the Jackson kernel and more generally $K^{(\ell)}(f)$ the ℓth order derivative. We will frequently use the second order derivative of the Jackson kernel evaluated at zero $K''(0)$, whose value is [16, Above Eq. (IV.5)]

$$K''(0) = -\frac{4\pi^2(M^2-1)}{3} = -\frac{\pi^2(n^2-4)}{3}.$$

Here we used the convention that $n = 2M$. Then its absolute value $|K''(0)|$ (denoted by τ) falls into the interval

$$|K''(0)| \in \left[\left(\frac{\pi^2}{3} - \frac{4\pi^2}{3(130)^2}\right) n^2, \left(\frac{\pi^2}{3}\right) n^2\right], \quad \text{for } n \geq 130.$$

For ease of exposition, we give an explicit lower bound on $|K''(0)|$ (which is valid for any $n \geq 130$):

$$\tau := |K''(0)| \geq \left(\frac{\pi^2}{3} - \frac{4\pi^2}{3(130)^2}\right) n^2 \geq 3.289 n^2, \quad \text{for } n \geq 130. \tag{A.2}$$

At a high-level, the purpose of introducing $\tau = |K''(0)|$ is to normalize the second order derivative of the Jackson kernel to 1 at $f = 0$.

A.1.1 Decomposing the Jackson Kernel

The Jackson kernel admits the following decomposition [16]

$$K(f_2-f_1)=\left[\frac{\sin(\pi M(f_2-f_1))}{M\sin(\pi(f_2-f_1))}\right]^4=\mathbf{a}(f_1)^H\mathbf{Z}\mathbf{a}(f_2)=\frac{1}{M}\sum_{j=-2M}^{2M}g_M(j)e^{-i2\pi j(f_2-f_1)},$$

where $M=n/2$ and $\mathbf{Z}$ is an $n\times n$ diagonal matrix whose diagonal entries are given by $[\mathbf{Z}]_{\ell\ell}=\frac{g_M(\ell)}{M}$ with

$$g_M(\ell)=\frac{1}{M}\sum_{k=\max(\ell-M,-M)}^{\min(\ell+M,M)}\left(1-\left|\frac{k}{M}\right|\right)\left(1-\left|\frac{\ell-k}{M}\right|\right)\geq 0,\ell=-2M,\ldots,2M, \tag{A.3}$$

the convolution of two discrete triangle functions scaled by $1/M$. The weighting function $g_M(\ell)$ attains its peak at zero and

$$g_M(0)=\frac{1}{M}\sum_{k=-M}^{M}\left(1-\left|\frac{k}{M}\right|\right)^2=\frac{2}{3}+\frac{2}{M^2}\overset{①}{\leq}\frac{2}{3}+\frac{2}{65^2}\leq 0.667,$$

where ① holds for $M\geq 65$ (or $n\geq 130$) by noting that $2/M^2$ is a decreasing function of M. Using the definition of $\mathbf{Z}$, we bound $\|\mathbf{Z}\|_{\infty,\infty}$ as

$$\|\mathbf{Z}\|_{\infty,\infty}=\max_{-2M\leq j\leq 2M}\frac{g_M(j)}{M}=\frac{g_M(0)}{M}\leq\frac{0.667}{M},\quad\text{for } n\geq 130. \tag{A.4}$$

A.1.2 Decomposing the Jackson Kernel Matrices

We frequently use matrices formed by sampling the Jackson kernel and its derivatives at appropriate frequencies. Given a finite set of frequencies $T=\{f_\ell\}_{\ell=1}^k$ (or its vector form $\mathbf{f}\in\mathbb{R}^k$), we define

$$\begin{aligned}
&\mathbf{D}_0(\mathbf{f}):=[K(f_m-f_n)]_{1\leq n\leq k,1\leq m\leq k}=\mathbf{A}(\mathbf{f})^H\mathbf{Z}\mathbf{A}(\mathbf{f});\\
&\mathbf{D}_1(\mathbf{f}):=[K'(f_m-f_n)]_{1\leq n\leq k,1\leq m\leq k}=\mathbf{A}(\mathbf{f})^H\mathbf{Z}\mathbf{A}'(\mathbf{f})=-\mathbf{A}'(\mathbf{f})^H\mathbf{Z}\mathbf{A}(\mathbf{f});\\
&\mathbf{D}_2(\mathbf{f}):=[K''(f_m-f_n)]_{1\leq n\leq k,1\leq m\leq k}=-\mathbf{A}'(\mathbf{f})^H\mathbf{Z}\mathbf{A}'(f)=\mathbf{A}''(\mathbf{f})^H\mathbf{Z}\mathbf{A}(\mathbf{f})=\mathbf{A}(\mathbf{f})^H\mathbf{Z}\mathbf{A}''(\mathbf{f}),
\end{aligned} \tag{A.5}$$

where

$$\mathbf{A}(\mathbf{f}):=[\mathbf{a}(f_1),\ldots,\mathbf{a}(f_k)],\ \mathbf{A}'(\mathbf{f}):=i2\pi\,\mathrm{diag}(\mathbf{n})\mathbf{A}(\mathbf{f}),\ A''(\mathbf{f}):=(i2\pi\,\mathrm{diag}(\mathbf{n}))^2\mathbf{A}(\mathbf{f})$$

with $\mathbf{n}=[-n,-n+1,\ldots,0,\ldots,n-1,n]^T$. More generally, the kernel matrix $\mathbf{D}_\ell(\mathbf{f}):=[K^{(\ell)}(f_m-f_n)]_{1\leq n\leq k,1\leq m\leq k}$ satisfies the factorization

$$\mathbf{D}_\ell(\mathbf{f})=(-1)^j\mathbf{A}^{(j)}(\mathbf{f})^H\mathbf{Z}\mathbf{A}^{(\ell-j)}(\mathbf{f}),\quad\text{for } j\leq\ell, \tag{A.6}$$

where $\mathbf{A}^{(\ell)}(\mathbf{f})$ represents the ℓth order derivative of the matrix $\mathbf{A}(\mathbf{f})$:

$$\mathbf{A}^{(\ell)}(\mathbf{f})=(i2\pi\,\mathrm{diag}(\mathbf{n}))^\ell\mathbf{A}(\mathbf{f}).$$

Similarly, we define the cross kernel matrices with respect to the frequency pair $(\mathbf{f}^1, \mathbf{f}^2)$ or $(\{f_\ell^1\}, \{f_\ell^2\})$ as

$$\mathbf{D}_\ell(\mathbf{f}^1, \mathbf{f}^2) = [K^{(\ell)}(f_m^2 - f_n^1)]_{1\le n\le k, 1\le m\le k}, \quad \text{for } \ell = 0, 1, 2.$$

We can also express $\mathbf{D}_\ell(\mathbf{f}^1, \mathbf{f}^2)$ in factorization forms

$$\mathbf{D}_\ell(\mathbf{f}^1, \mathbf{f}^2) = (-1)^j \mathbf{A}^{(j)}(\mathbf{f}^1)^H \mathbf{Z} \mathbf{A}^{(\ell-j)}(\mathbf{f}^2), \quad \text{for } j \le \ell. \tag{A.7}$$

A.1.3 Bounding the Jackson Kernel

The following lemma provides a set of bounds on the ℓth derivative of the Jackson kernel for $\ell \in \{0, 1, 2, 3, 4\}$.

Lemma A.1.1 (Bounds on $|K^{(\ell)}|$)**.** *For $\ell \in \{0, 1, 2, 3, 4\}$, let $K^{(\ell)}$ be the ℓth derivative of K ($K = K^{(0)}$). Define $s(f)$ as a symmetric and periodic function with period 1 and $s(f) = \frac{1}{Mf(3-4f^2)}$ for $f \in (0, 1/2]$. Then for $f \in (0, 1/2]$, we have*

$$\begin{aligned}
&|K^{(0)}(f)| \le B_0(f) := s^4(f), \\
&|K^{(1)}(f)| \le B_1(f) := 2\pi M s^4(f) \left(\frac{3\sqrt{3}}{8} + 2s(f) \right), \\
&|K^{(2)}(f)| \le B_2(f) := (2\pi M)^2 s^4(f) \left(1 + \frac{3\sqrt{3}}{2} s(f) + 5s^2(f) \right), \\
&|K^{(3)}(f)| \le B_3(f) := (2\pi M)^3 s^4(f) \left(c_1 + 6s(f) + \frac{45\sqrt{3}}{8} s^2(f) + 15s^3(f) \right), \\
&|K^{(4)}(f)| \le B_4(f) := (2\pi M)^4 s^4(f) \left(\frac{5}{2} + c_2 s(f) + 30s^2(f) + \frac{45\sqrt{3}}{2} s^3(f) + \frac{105}{2} s^4(f) \right),
\end{aligned}$$

where

$$\begin{aligned}
c_1 &= \frac{1}{2} \left(\sin\left(2\tan^{-1}\left(\sqrt{\frac{1}{5}\left(\sqrt{129}+12\right)} \right) \right) - 2\sin\left(4\tan^{-1}\left(\sqrt{\frac{1}{5}\left(\sqrt{129}+12\right)} \right) \right) \right), \\
c_2 &= -4\sin\left(2\tan^{-1}\left(\sqrt{\frac{1}{5}\left(\sqrt{129}+12\right)} \right) \right) \left(4\cos\left(2\tan^{-1}\left(\sqrt{\frac{1}{5}\left(\sqrt{129}+12\right)} \right) \right) - 1 \right).
\end{aligned}$$

Furthermore, $B_\ell(f)$ is decreasing in f on $(0, 1/2]$ and $B_\ell(\Omega - f) + B_\ell(\Omega + f)$ is increasing in f for any positive Ω such that $\Omega > f$ and $\Omega + f \le 1/2$.

Proof. We need the following elementary bound on the sine function for $f \in [0, \frac{1}{2}]$:

$$\sin(\pi f) \ge f(3 - 4f^2). \tag{A.8}$$

Clearly, a consequence is $\frac{1}{M|\sin(\pi f)|} \le s(f), f \in [-\frac{1}{2}, \frac{1}{2}] \backslash \{0\}$. We use this fact together with explicit expressions for $K^{(\ell)}(f)$ to develop upper bounds.

When $\ell = 0$,

$$|K(f)| = \left| \frac{\sin(\pi M f)}{M \sin(\pi f)} \right|^4 \leq s^4(f).$$

When $\ell = 1$,

$$K^{(1)}(f) = \frac{2\pi M}{(M \sin(\pi f))^4} \frac{1}{M} \left(-2 \cot(\pi f) \sin^4(\pi f M) + 2 \sin^3(\pi f M) \cos(\pi f M) M \right)$$

implying

$$|K^{(1)}(f)| \leq 2\pi M s^4(f) \left(\frac{3\sqrt{3}}{8} + 2s(f) \right),$$

since $\max_f |2 \cos(\pi f M) \sin(\pi f M)^3| \leq \frac{3\sqrt{3}}{8}$.

When $\ell = 2$,

$$K^{(2)}(f) = \frac{(2\pi M)^2}{(M \sin(\pi f))^4} \frac{1}{M^2} \times$$
$$\left((2 \cos(2\pi f) + 3) \csc^2(\pi f) \sin^4(\pi f M) - 8 \cot(\pi f) \sin^3(\pi f M) \cos(\pi f M) M + \sin^2(\pi f M)(2 \cos(2\pi f M) + 1) M^2 \right)$$

implying

$$|K^{(2)}(f)| \leq (2\pi M)^2 s^4(f) \left(1 + \frac{3\sqrt{3}}{2} s(f) + 5s^2(f) \right),$$

where we used $\max_f |8 \sin^3(\pi f M) \cos(\pi f M)| = \frac{3\sqrt{3}}{2}$ and $\max_f |\sin^2(\pi f M)(2 \cos(2\pi f M) + 1)| = 1$.

When $\ell = 3$,

$$K^{(3)}(f) = \frac{(2\pi M)^3}{(M \sin(\pi f))^4} \frac{1}{M^3} \times \Bigg(- (4 \cos(2\pi f) + 11) \cot(\pi f) \csc^2(\pi f) \sin^4(\pi f M)$$
$$+ 6(2 \cos(2\pi f) + 3) \csc^2(\pi f) \sin^3(\pi f M) \cos(\pi f M) M$$
$$- 6 \cot(\pi f) \sin^2(\pi f M)(2 \cos(2\pi f M) + 1) \sin(4\pi f M) M^2 - \frac{1}{2} \sin(2\pi f M) M^3 \Bigg)$$

implying

$$|K^{(3)}(f)| \leq (2\pi M)^3 s^4(f) \left(c_1 + 6s(f) + \frac{45\sqrt{3}}{8} s^2(f) + 15s^3(f) \right),$$

by recognizing the following upper bounds:

$$\max_{f \in [0,1/2]} |(4 \cos(2\pi f) + 11) \cos(\pi f)| = 15, \quad \max_f |6 \sin^2(\pi f M)(2 \cos(2\pi f M) + 1)| = 6,$$
$$\max_{f \in [0,1/2]} |6(2 \cos(2\pi f) + 3)| = 30\,, \quad \max_f |\sin(4\pi f M) - (1/2) \sin(2\pi f M)| = c_1,$$
$$\max_f |\sin^3(\pi f M) \cos(\pi f M)| = 3\sqrt{3}/16.$$

When $\ell = 4$,

$$K^{(4)}(f) = \frac{(2\pi M)^4}{(M\sin(\pi f))^4}\frac{1}{M^4} \times \Bigg(\frac{1}{2}(49\cos(2\pi f) + 4\cos(4\pi f) + 52)\csc^4(\pi f)\sin^4(\pi f M) \\ - 8(4\cos(2\pi f) + 11)\cot(\pi f)\csc^2(\pi f)\sin^3(\pi f M)\cos(\pi f M)M \\ + 6(2\cos(2\pi f) + 3)\csc^2(\pi f)\sin^2(\pi f M)(2\cos(2\pi f M) + 1)M^2 \\ - 4\cot(\pi f)\sin(2\pi f M)(4\cos(2\pi f M) - 1)M^3 \\ + (2\cos(4\pi f M) - \frac{1}{2}\cos(2\pi f M))M^4\Bigg)$$

implying

$$|K^{(4)}(f)| \le (2\pi M)^4 s^4(f)\left(\frac{5}{2} + c_2 s(f) + 30s^2(f) + \frac{45\sqrt{3}}{2}s^3(f) + \frac{105}{2}s^4(f)\right),$$

which follows from the following upper bounds:

$\max_{f\in[0,1/2]} \frac{1}{2}(49\cos(2\pi f) + 4\cos(4\pi f) + 52) = 105/2$, $\max_f |\sin^2(\pi f M)(2\cos(2\pi f M) + 1)| = 1$,
$\max_{f\in[0,1/2]} |8(4\cos(2\pi f) + 11)\cos(\pi f)| = 120$, $\max_f |4\sin(2\pi f M)(4\cos(2\pi f M) - 1)| = c_2$,
$\max_f |\sin^3(\pi f M)\cos(\pi f M)| = 3\sqrt{3}/16$, $\max_f |2\cos(4\pi f M) - 1/2\cos(2\pi f M)| = 5/2$,
$\max_{f\in[0,1/2]} |6(2\cos(2\pi f) + 3)| = 30$.

Finally, $s(f)$ is nonnegative and is decreasing in $(0, 1/2]$ since $s'(f)$ is negative on $(0, 1/2)$. Therefore, the kth power $s^k(f)$ is decreasing in $(0, 1/2])$, which further implies that $B_\ell(f), \ell = 0, 1, 2, 3, 4$ is decreasing in $(0, 1/2]$. In addition, since $s(f)$ is convex in $(0, 1/2]$, $s^k(f)$ is also convex as a consequence of the composition rule of convex and monotonic functions. Combining the convex and decreasing property of $s^k(f)$ on $(0, 1/2]$ and then applying arguments similar to those in [13, Lemma 2.6], we conclude that $B_\ell(\Omega - f) + B_\ell(\Omega + f)$ is increasing in f for any positive Ω such that $\Omega > f$ and $\Omega + f \le 1/2$.

□

A.1.4 Bounding the Sums of the Jackson Kernel

Without loss of generality, we assume $0 \in T$ and develop bounds on $\sum_{f_i\in T\setminus\{0\}} |K^{(\ell)}(f - f_i)|, \ell \in \{0, 1, 2, 3, 4\}$ when f lives in a neighborhood around 0. It is easy to verify the following lemma based on the properties of $|K^{(\ell)}(f)|, \ell = 0, 1, 2, 3, 4$ in Lemma A.1.1. The proof parallels that of [13, Lemma 2.7] and is omitted here.

Lemma A.1.2. *Suppose $0 \in T$ and f_+ is the smallest positive frequency in T. Let $\Delta := \Delta(T) \ge \Delta_{\min}$ and $f \in [0, \bar{f}]$ where $\bar{f} \le f_+/2$. Then for $\ell \in \{0, 1, 2, 3, 4\}$,*

$$\sum_{f_i\in T\setminus\{0\}} |K^{(\ell)}(f - f_i)| \le F_\ell(\Delta, f) := F_\ell^+(\Delta, f) + F_\ell^-(\Delta, f) \le F_\ell(\Delta_{\min}, \bar{f})$$

with

$$F_\ell^+(\Delta, f) = \max\left\{\max_{\Delta\le\xi\le 3\Delta_{\min}} |K^{(\ell)}(f-\xi)|, B_\ell(3\Delta_{\min}-f)\right\} + \sum_{j=2}^{\lfloor\frac{1}{2\Delta_{\min}}\rfloor} B_\ell(j\Delta_{\min}-f),$$

$$F_\ell^-(\Delta, f) = \max\left\{\max_{\Delta\le\xi\le 3\Delta_{\min}} |K^{(\ell)}(\xi)|, B_\ell(3\Delta_{\min})\right\} + \sum_{j=2}^{\lfloor\frac{1}{2\Delta_{\min}}\rfloor} B_\ell(j\Delta_{\min}+f).$$

$F_\ell(\Delta, f)$ is decreasing in Δ. When Δ is fixed as $\Delta_{\min}$, $F_\ell(\Delta_{\min}, f)$ is increasing in f.

The following lemma provides bounds on $\sum_{f_i\in T} |K^{(\ell)}(f-f_i)|$ for $\ell \in \{0,1,2,3,4\}$ and is a direct consequence of the decreasing property of $B_\ell(\cdot)$.

Lemma A.1.3. *Suppose $0 \in T$, f_+ is the smallest positive frequency in T and $f \in [\underline{f}, f_+ - \bar{f}]$. Then for $\ell \in \{0,1,2,3,4\}$,*

$$\sum_{f_i\in T} |K^{(\ell)}(f-f_i)| \le W_\ell(\underline{f}, \bar{f}) := \sum_{j=0}^{\lfloor\frac{1}{2\Delta_{\min}}\rfloor} B_\ell(j\Delta_{\min}+\underline{f}) + \sum_{j=0}^{\lfloor\frac{1}{2\Delta_{\min}}\rfloor} B_\ell(j\Delta_{\min}+\bar{f}).$$

A.1.5 Numerical Bounds on the Jackson Kernel Sums

Suppose $0 \in T$. Then by Lemma A.1.2 we can bound $\sum_{f_j\in T\setminus\{0\}} |K^{(\ell)}(f-f_j)|$ for $f \in [0, \bar{f}]$ as:

$$\sum_{f_j\in T\setminus\{0\}} |K^{(\ell)}(f-f_j)| \le F_\ell(\Delta_{\min}, \bar{f}).$$

We list the values of $F_\ell(\Delta_{\min}, \bar{f})$ for different $\bar{f}$ in Table A.1.

We can use Lemma A.1.3 to bound $\sum_{f_j\in T} |K^{(\ell)}(f-f_j)|$ for $f \in [\underline{f}, f_+ - \bar{f}]$ as

$$\sum_{f_j\in T} |K^{(\ell)}(f-f_j)| \le W_\ell(\underline{f}, \bar{f}).$$

We list the values of $W_\ell(\underline{f}, \bar{f})$ for different $\underline{f}, \bar{f}$ in Table A.2.

Finally, we list several numerical upper bounds on $|K^{(\ell)}(f)|$ and $K''(f)$ over different intervals in Table A.3, which directly follow from [13, equations (2.21)-(2.24)] and numerical computations.

A.1.6 Controlling the Jackson Kernel Matrices

In this section, we derive several consequences of the joint frequency-coefficient vector $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{u}, \mathbf{v})$ living in the neighborhood $\mathcal{N}^\star$ of the true joint frequency-coefficient vector $\boldsymbol{\theta}^\star = (\mathbf{f}^\star, \mathbf{u}^\star, \mathbf{v}^\star)$. Recall that $\mathcal{N}^\star$ contains all $\boldsymbol{\theta}$ that is close to $\boldsymbol{\theta}^\star$ in the ℓ_∞ norm:

Table A.1: Numerical upper bounds on $F_\ell(2.5/n, f)$.

f	$F_0(2.5/n, f)$	$F_1(2.5/n, f)$	$F_2(2.5/n, f)$	$F_3(2.5/n, f)$	$F_4(2.5/n, f)$
0	0.00755	$0.01236n$	$0.05610n^2$	$0.28687n^3$	$1.48634n^4$
$0.002/n$	0.00755	$0.01236n$	$0.05610n^2$	$0.28687n^3$	$1.48634n^4$
$0.24/n$	0.00757	$0.01241n$	$0.05637n^2$	$0.28838n^3$	$1.67097n^4$
$0.2404/n$	0.00757	$0.01241n$	$0.05637n^2$	$0.28838n^3$	$1.67100n^4$
$0.75/n$	0.00772	$0.01450n$	$0.12639n^2$	$1.07987n^3$	$6.57069n^4$
$0.7504/n$	0.00772	$0.01454n$	$0.12675n^2$	$1.08211n^3$	$6.57595n^4$

Table A.2: Numerical upper bounds on $W_\ell(f_1, f_2)$.

f_1	f_2	$W_0(f_1, f_2)$	$W_1(f_1, f_2)$	$W_2(f_1, f_2)$
$0.7496/n$	$1.25/n$	0.71059	$5.2265n$	$48.0330n^2$
$0.75/n$	$1.25/n$	0.70859	$5.2084n$	$47.8388n^2$

Table A.3: Numerical upper bounds on $|K^{(\ell)}(f)|$ and $K''(f)$.

f	$\lvert K(f)\rvert$	$\lvert K'(f)\rvert$	$\lvert K''(f)\rvert$	$\lvert K'''(f)\rvert$	$\lvert K''''(f)\rvert$	$K''(f)$
$[0, 0.002/n]$	1	$0.00658n$	$3.290n^2$	$0.0649394n^3$		
$[0, 0.24/n]$	1	$0.789569n$	$3.290n^2$	$7.79273n^3$		$-2.35084n^2$
$[0, 0.2404/n]$	1	$0.790885n$	$3.290n^2$	$7.80572n^3$	$29.2227n^4$	
$[0.2396/n, 0.7504/n]$	0.90951	$2.46872n$	$3.290n^2$			

$$\mathcal{N}^\star = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_{\hat{\infty}} \le X^\star \gamma_0/\sqrt{2}\}. \tag{A.9}$$

Recall that the weighted ℓ_∞ norm $\|\cdot\|_{\hat{\infty}}$ is defined by $\|(\mathbf{f}, \mathbf{u}, \mathbf{v})\|_{\hat{\infty}} := \|(\mathbf{Sf}, \mathbf{u}, \mathbf{v})\|_\infty$ with $\mathbf{S} := \sqrt{\tau}\,\mathrm{diag}(|\mathbf{c}^\star|)$. We remark that all the results in this section still hold for the bigger neighborhood $\hat{\mathcal{N}}$ defined by replacing $X^\star$ with $\hat{X} = X^\star + 35.2$. Indeed, for the results to hold, the key requirement on $\boldsymbol{\theta}$ is $\|\mathbf{f} - \mathbf{f}^\star\|_\infty \le 0.002/n$. This condition holds for both regions because as we will show later

$$\|\mathbf{f} - \mathbf{f}^\star\|_\infty \le \begin{cases} 0.4X^\star\gamma & \text{for } \boldsymbol{\theta} \in \mathcal{N}^\star, \\ 0.4\hat{X}\gamma & \text{for } \boldsymbol{\theta} \in \hat{\mathcal{N}}. \end{cases}$$

Invoking the SNR condition (2.10), we conclude that the two upper bounds are much smaller than $0.002/n$ in both cases.

Our first set of results bound the distances between the parameters in $\boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}$: for each $j = 1, \ldots, k$

$$\begin{aligned} \frac{|c_j - c_j^\star|}{|c_j^\star|} &\overset{①}{\leq} X^\star\gamma, \\ \frac{|c_j|}{|c_j^\star|} &\overset{②}{\leq} 1 + X^\star\gamma, \\ \left|\left(|c_j|/|c_j^\star|\right)^2 - 1\right| &\overset{③}{\leq} X^\star\gamma(2 + X^\star\gamma), \\ |f_j - f_j^\star| &\overset{④}{\leq} X^\star\gamma/\sqrt{2\tau} \overset{⑤}{\leq} 0.4X^\star\gamma/n. \end{aligned} \tag{A.10}$$

For ① to hold, first note $\|\mathbf{u} - \mathbf{u}^\star\|_\infty \leq X^\star\gamma_0/\sqrt{2}$ and $\|\mathbf{v} - \mathbf{v}^\star\|_\infty \leq X^\star\gamma_0/\sqrt{2}$ by (A.9). Also note

$$\begin{aligned} \|\mathbf{c} - \mathbf{c}^\star\|_\infty^2 &= \max_\ell[(u_\ell - u_\ell^\star)^2 + (v_\ell - v_\ell^\star)^2] \\ &\leq \max_\ell(u_\ell - u_\ell^\star)^2 + \max_\ell(v_\ell - v_\ell^\star)^2 \\ &= \|\mathbf{u} - \mathbf{u}^\star\|_\infty^2 + \|\mathbf{v} - \mathbf{v}^\star\|_\infty^2 \leq 2(X^\star\gamma_0/\sqrt{2})^2 = (X^\star\gamma_0)^2. \end{aligned}$$

Finally ① follows since $\max_j |c_j - c_j^\star|/|c_j^\star| \leq \|\mathbf{c} - \mathbf{c}^\star\|_\infty/c_{\min}^\star$ and $\gamma = \gamma_0/c_{\min}^\star$. After we show ①, ② follows from $|c_j|/|c_j^\star| = |c_j - c_j^\star + c_j^\star|/|c_j^\star|$ and the triangle inequality. ③ follows from $|(|c_j|/|c_j^\star|)^2 - 1| = |(|c_j|/|c_j^\star| + 1)(|c_j|/|c_j^\star| - 1)|$. ④ follows from the definition of the $\ell_{\hat{\infty}}$ norm:

$$\begin{aligned} &\|S(\mathbf{f} - \mathbf{f}^\star)\|_\infty \leq X^\star\gamma_0/\sqrt{2} \\ \Longrightarrow &\|\sqrt{\tau}\,\mathrm{diag}(|\mathbf{c}^\star|)(\mathbf{f} - \mathbf{f}^\star)\|_\infty \leq X^\star\gamma_0/\sqrt{2} \\ \Longrightarrow &|f_j - f_j^\star| \leq X^\star\gamma_0/|c_j^\star|/\sqrt{2\tau},\ \forall j \\ \Longrightarrow &|f_j - f_j^\star| \leq X^\star\gamma_0/c_{\min}^\star/\sqrt{2\tau} = X^\star\gamma/\sqrt{2\tau},\ \forall j. \end{aligned}$$

Finally ⑤ holds due to the fact that $\tau \geq 3.289n^2$ for $n \geq 130$ by (A.2) and hence

$$1/\sqrt{2\tau} \leq 1/\sqrt{2(3.289)}/n \leq 0.3899/n \leq 0.4/n.$$

Next, we present the second class of results that quantify the well-conditionedness of the Jackson kernel matrices $\mathbf{D}_\ell(\mathbf{f}),\ \ell = 0, 1, 2$. Such results are instrumental to dual certificate construction [13]. Since the minimal separation $\Delta(T)$ is a key quantity affecting the well-conditionedness, we first show that those frequencies $T^\lambda := \{f_\ell^\lambda\}$ and $\hat{T} := \{\hat{f}_\ell\}$ in Lemma 2.4.1 and Lemma 2.4.2 satisfy a separation condition, provided $T^\star = \{f_\ell^\star\}$ satisfy a slightly stronger separation condition. The proof is given in Appendix A.11.

Lemma A.1.4. *Let the separation condition* (2.9) *and the SNR condition* (2.10) *hold. Then both the frequencies* $T^\lambda = \{f_\ell^\lambda\}$ *returned by the first fixed point map* (2.19) *and the frequencies* $\hat{T} = \{\hat{f}_\ell\}$ *generated by the second fixed point map* (2.21) *have minimal separations at least* $2.5/n$*. Furthermore, the intermediate frequencies defined by* $\tilde{T} = \{\tilde{f}_\ell\}_{\ell=1}^k$ *with each* $\tilde{f}_\ell \in [f_\ell^\star, f_\ell^\lambda]$ *or* $[f_\ell^\lambda, f_\ell^\star]$ *and the second intermediate frequencies* $\tilde{T}^\lambda := \{\tilde{f}_\ell\}_{\ell=1}^k$ *with each*

$\tilde{f}_\ell \in [f_\ell^\lambda, \hat{f}_\ell]$ *or* $[\hat{f}_\ell, f_\ell^\lambda]$ *also have minimal separations at least* $2.5/n$*:*

$$\min\{\Delta(T^\lambda), \Delta(\tilde{T}), \Delta(\hat{T}), \Delta(\tilde{T}^\lambda)\} \geq 2.5/n.$$

Now we are ready to provide numerical bounds related to the well-conditionedness of the Jackson kernel matrices $\mathbf{D}_\ell(\mathbf{f}),\ \ell = 0, 1, 2$:

$$\begin{aligned} \|\mathbf{I} - \mathbf{D}_0(\mathbf{f})\|_{\infty,\infty} &\overset{①}{\leq} F_0(2.5/n, 0) \overset{④}{\leq} 0.00755, \\ \left\|\mathbf{D}_1(\mathbf{f})/\sqrt{\tau}\right\|_{\infty,\infty} &\overset{②}{\leq} F_1(2.5/n, 0)/\sqrt{\tau} \overset{⑤}{\leq} 0.01236n/\sqrt{\tau} \leq 0.00682, \\ \|\mathbf{I} - (-\mathbf{D}_2(\mathbf{f})/\tau)\|_{\infty,\infty} &\overset{③}{\leq} F_2(2.5/n, 0)/\tau \overset{⑥}{\leq} 0.05610n^2/\tau \leq 0.0171, \end{aligned} \tag{A.11}$$

where ①, ② and ③ follow because the diagonal entries of these kernel matrices are given by $[\mathbf{D}_0(\mathbf{f}, \mathbf{f})]_{\ell,\ell} = K(0) = 1$, $[\mathbf{D}_1(\mathbf{f}, \mathbf{f})]_{\ell,\ell} = K'(0) = 0$ and $[\mathbf{D}_2(\mathbf{f}, \mathbf{f})]_{\ell,\ell} = K''(0) = -\tau$ [16, Section IV.A]. Hence, it suffices to compute $\sum_{f_i \in T\setminus\{\zeta\}} |K^{(\ell)}(\zeta - f_i)|$ for $\zeta \in T$ which can be bounded by $F_\ell(2.5/n, 0)$ according to Lemma A.1.2 since $\Delta(T) \geq 2.5/n$ by Lemma A.1.4. The inequalities ④, ⑤ and ⑥ follow from the upper bounds on $F_\ell(2.5/n, 0)$ in Table A.1 and the fact that $\tau \geq 3.289n^2$ for $n \geq 130$ by (A.2).

To control the $\ell_{\infty,\infty}$ distance between two kernel matrices, say $\mathbf{D}_0(\mathbf{f})$ and $D_0(\mathbf{f}, f^\star)$, we apply the mean value theorem and Lemma A.1.2:

$$\begin{aligned} \|\mathbf{D}_0(\mathbf{f}) - \mathbf{D}_0(\mathbf{f}, \mathbf{f}^\star)\|_{\infty,\infty} &\overset{①}{=} \|\mathbf{D}_0(f_1, \mathbf{f}) - \mathbf{D}_0(f_1, \mathbf{f}^\star)\|_1 \\ &\leq \sum_\ell |K(f_\ell - f_1) - K(f_\ell^\star - f_1)| \\ &\overset{②}{=} \sum_\ell |K'(\tilde{f}_\ell - f_1)(f_\ell - f_\ell^\star)| \\ &\leq (|K'(\tilde{f}_1 - f_1)| + \sum_{\ell \neq 1} |K'(\tilde{f}_\ell - f_1)|)\|\mathbf{f} - \mathbf{f}^\star\|_\infty \\ &\overset{③}{\leq} (F_1(2.5/n, 0.002/n) + \max_{f \in [0, 0.002/n]} |K'(f)|)\|\mathbf{f} - \mathbf{f}^\star\|_\infty \\ &\overset{④}{\leq} (0.01236n + 0.00658n)(0.4X^\star\gamma/n) = 0.00758X^\star\gamma, \end{aligned} \tag{A.12}$$

where ① follows since by rearranging indices if necessary, we can assume without loss of generality that the maximum absolute row sum of $\mathbf{D}_0(\mathbf{f}) - \mathbf{D}_0(\mathbf{f}, \mathbf{f}^\star)$ happens at the first row; ② holds because we applied the mean value theorem for some $\tilde{f}_\ell$ between f_ℓ and $f_\ell^\star$; ③ follows from the monotonic property of $F_\ell(2.5/n, f)$ in Lemma A.1.2 by taking into account that $\Delta(\tilde{T}) \geq 2.5/n$ (by Lemma A.1.4) and $\|\tilde{\mathbf{f}} - \mathbf{f}\|_\infty \leq \|\mathbf{f}^\star - \mathbf{f}\|_\infty \leq 0.4X^\star\gamma/n \leq 0.002/n$. ④ follows from the upper bounds on $F_1(2.5/n, 0.002/n)$ in Table A.1 and $\max_{f \in [0, 0.002/n]} |K'(f)|$ in Table A.3.

Applying the similar arguments as the step ③, we can get a more general result as follows

Lemma A.1.5. *Let an arbitrary cluster of points* $T := \{f_j\}$ *satisfy the separation condition of* $\Delta(T) \geq 2.5/n$*. Assume* $\underline{f} \leq |f - f_r| \leq \bar{f}$ *for an arbitrary* $f_r \in T$*. Then,*

$$\sum_j |K^{(\ell)}(f_j - f)| \le F_\ell(2.5/n, \bar{f}) + \max_{f\in[\underline{f},\bar{f}]} |K^{(\ell)}(f)|. \tag{A.13}$$

To control $\|\mathbf{D}_\ell(\mathbf{f},\mathbf{f}^\star) - \mathbf{D}_\ell(\mathbf{f})\|_{\infty,\infty}$ in a similar manner for $\ell = 1, 2$, we note that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_\infty \le X^\star\gamma_0/\sqrt{2}$ and both T and $\tilde{T}$ are well-separated: $\Delta(T) \ge 2.5/n$ and $\Delta(\tilde{T}) \ge 2.5/n$ with $\tilde{T}$ composed of certain "middle" frequencies $\tilde{f}_\ell \in [f_\ell, f_\ell^\star]$ or $[f_\ell^\star, f_\ell]$. Then using Lemma A.1.5, we upperbound $\|\mathbf{D}_\ell(\mathbf{f},\mathbf{f}^\star) - \mathbf{D}_\ell(\mathbf{f})\|_{\infty,\infty}$ as follows

$$\begin{aligned}\frac{1}{\sqrt{\tau}}\|\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star) - \mathbf{D}_1(\mathbf{f})\|_{\infty,\infty} &\overset{①}{\le} 1/\sqrt{\tau}\,(F_2(2.5/n, 0.002/n) + \max_{f\in[0,0.002/n]} |K''(f)|)\|\mathbf{f} - \mathbf{f}^\star\|_\infty \\ &\overset{②}{\le} (1/\sqrt{3.289n^2})(0.05610n^2 + 3.290n^2)(0.4X^\star\gamma/n) \le 0.73802X^\star\gamma,\end{aligned} \tag{A.14}$$

where ① follows by Lemma A.1.5 and ② follows from the fact that $\tau \ge 3.289n^2$ for $n \ge 130$ in (A.2) and by combining the upper bound on $F_2(2.5/n, 0.002/n)$ in Table A.1 and the upper bound on $\max_{f\in[0,0.002/n]} |K''(f)|$ in Table A.3. Similarly, following from Lemma A.1.5 and the mean value theorem, by combining the upper bound on $F_3(2.5/n, 0.002/n)$ in Table A.1 and the upper bound on $\max_{f\in[0,0.002/n]} |K^{(3)}(f)|$ in Table A.3, we have

$$\begin{aligned}\frac{1}{\tau}\|\mathbf{D}_2(\mathbf{f},\mathbf{f}^\star) - \mathbf{D}_2(\mathbf{f})\|_{\infty,\infty} &\le \frac{1}{\tau}(F_3(2.5/n, 0.002/n) + \max_{f\in[0,0.002/n]} |K'''(f)|)\|\mathbf{f} - \mathbf{f}^\star\|_\infty \\ &\le (1/3.289n^2)(0.28687n^3 + 0.0649394n^3)(0.4X^\star\gamma/n) = 0.04279X^\star\gamma.\end{aligned} \tag{A.15}$$

To control $\|\mathbf{D}_\ell(\mathbf{f}^\star) - \mathbf{D}_\ell(\mathbf{f})\|_{\infty,\infty}$, we rewrite $\mathbf{D}_\ell(\mathbf{f}^\star) - \mathbf{D}_\ell(\mathbf{f})$ as

$$\mathbf{D}_\ell(\mathbf{f}^\star) - \mathbf{D}_\ell(\mathbf{f}) = \mathbf{D}_\ell(\mathbf{f}^\star) - \mathbf{D}_\ell(\mathbf{f}^\star,\mathbf{f}) + \mathbf{D}_\ell(\mathbf{f}^\star,\mathbf{f}) - \mathbf{D}_\ell(\mathbf{f}).$$

Then, the desired results follow from the triangle inequality of the $\ell_{\infty,\infty}$ norm:

$$\begin{aligned}\|\mathbf{D}_0(\mathbf{f}^\star) - \mathbf{D}_0(\mathbf{f})\|_{\infty,\infty} &\le \|\mathbf{D}_0(\mathbf{f}^\star) - \mathbf{D}_0(\mathbf{f}^\star,\mathbf{f})\|_{\infty,\infty} + \|\mathbf{D}_0(\mathbf{f}^\star,\mathbf{f}) - \mathbf{D}_0(\mathbf{f})\|_{\infty,\infty} \\ &\overset{①}{\le} 2(0.00758X^\star\gamma) = 0.01516X^\star\gamma,\end{aligned} \tag{A.16}$$

where ① follows from (A.12) and an exchange of the roles of $\mathbf{f}$ and $\mathbf{f}^\star$;

$$\begin{aligned}\frac{1}{\sqrt{\tau}}\|\mathbf{D}_1(\mathbf{f}^\star) - \mathbf{D}_1(\mathbf{f})\|_{\infty,\infty} &\le \frac{1}{\sqrt{\tau}}\|\mathbf{D}_1(\mathbf{f}^\star) - \mathbf{D}_1(\mathbf{f}^\star,\mathbf{f})\|_{\infty,\infty} + \frac{1}{\sqrt{\tau}}\|\mathbf{D}_1(\mathbf{f}^\star,\mathbf{f}) - \mathbf{D}_1(\mathbf{f})\|_{\infty,\infty} \\ &\overset{①}{\le} 2(0.73802X^\star\gamma) = 1.47604X^\star\gamma,\end{aligned} \tag{A.17}$$

where ① follows from (A.14);

$$\begin{aligned}\frac{1}{\tau}\|\mathbf{D}_2(\mathbf{f}^\star) - \mathbf{D}_2(\mathbf{f})\|_{\infty,\infty} &\le \frac{1}{\tau}\|\mathbf{D}_2(\mathbf{f}^\star) - \mathbf{D}_2(\mathbf{f}^\star,\mathbf{f})\|_{\infty,\infty} + \frac{1}{\tau}\|\mathbf{D}_2(\mathbf{f}^\star,\mathbf{f}) - \mathbf{D}_2(\mathbf{f})\|_{\infty,\infty} \\ &\overset{①}{\le} 2(0.04279X^\star\gamma) = 0.08558X^\star\gamma,\end{aligned} \tag{A.18}$$

where ① follows from (A.15).

Then following from Eq. (A.10), (A.14)- (A.15) and (A.16)- (A.18), and together with the sub-multiplicative property of the $\ell_{\infty,\infty}$ norm, we have

$$\begin{aligned}\left\|\frac{1}{\sqrt{\tau}}\operatorname{diag}(1./|\mathbf{c}^\star|)[\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star-\mathbf{D}_1(\mathbf{f})\mathbf{c}]\right\|_\infty \leq& \frac{1}{\sqrt{\tau}}\|\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)-\mathbf{D}_1(\mathbf{f})\|_{\infty,\infty}\|1./\mathbf{c}^\star\|_\infty\|\mathbf{c}^\star\|_\infty\\ &+\frac{1}{\sqrt{\tau}}\|\mathbf{D}_1(\mathbf{f})\|_{\infty,\infty}\|1./\mathbf{c}^\star\|_\infty\|\mathbf{c}-\mathbf{c}^\star\|_\infty\\ \leq&(0.73802X^\star\gamma)B^\star+(0.01236)B^\star(X^\star\gamma)=0.75038B^\star X^\star\gamma,\end{aligned} \tag{A.19}$$

where the last but one line follows from $\|1./\mathbf{c}^\star\|_\infty \leq 1/c^\star_{\min}$ and $\gamma = \gamma_0/c^\star_{\min}$. Here and throughout the rest of the paper, we use $1./\mathbf{x}$, $1./|\mathbf{x}|$, $\mathbf{y}./\mathbf{x}$, $|\mathbf{y}|./|\mathbf{x}|$, $\mathbf{x}\odot\mathbf{y}$ and $\frac{1}{\mathbf{x}}$, $\frac{1}{|\mathbf{x}|}$, $\frac{\mathbf{y}}{\mathbf{x}}$, $\frac{|\mathbf{y}|}{|\mathbf{x}|}$ in the sense of pointwise arithmetic operations, here $\mathbf{x},\mathbf{y}$ stand for any vectors of the same length.

We apply similar arguments to develop the following bound

$$\begin{aligned}\left\|\frac{1}{\tau}\operatorname{diag}(1./|\mathbf{c}^\star|)[\mathbf{D}_2(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star-\mathbf{D}_2(\mathbf{f})\mathbf{c}]\right\|_\infty \leq& \frac{1}{\tau}\|\mathbf{D}_2(\mathbf{f},\mathbf{f}^\star)-\mathbf{D}_2(\mathbf{f})\|_{\infty,\infty}\|1./\mathbf{c}^\star\|_\infty\|\mathbf{c}^\star\|_\infty\\ &+\frac{1}{\tau}\|\mathbf{D}_2(\mathbf{f})\|_{\infty,\infty}\|1./\mathbf{c}^\star\|_\infty\|\mathbf{c}-\mathbf{c}^\star\|_\infty\\ \leq&(0.08558X^\star\gamma)B^\star+(1.05610)B^\star(X^\star\gamma)\leq 1.14168B^\star X^\star\gamma.\end{aligned} \tag{A.20}$$

A.2 Bounding the Dual Atomic Norm of Gaussian Noise

In this section, we develop an upper bound on the dual atomic norm of the weighted Gaussian noise $\mathbf{Zw}\sim\mathcal{N}(0,\sigma^2Z^2)$ for the positive definite diagonal matrix $\mathbf{Z}$ with $[\mathbf{Z}]_{\ell,\ell}=\frac{g_M(\ell)}{M}$. First following [22, C.4 with $N\geq 4\pi(2n+1)$], we get

$$\sup_{f\in\mathbb{T}}\left|\mathbf{a}(f)^H\mathbf{Zw}\right|\leq 2\max_{m=0,\ldots,N-1}|S_m|, \tag{A.21}$$

where $\{S_m\}_{m=0}^{N-1}$ are N equispaced samples of the continuous function $\mathbf{a}(f)^H\mathbf{Zw}$ defined on $\mathbb{T}=[0,1]$:

$$S_m:=\mathbf{a}(\frac{m}{N})^H\mathbf{Zw}=\sum_{\ell=-n}^{n}\frac{g_M(\ell)}{M}w_\ell e^{-i2\pi\ell\frac{m}{N}}.$$

Since $\{w_\ell\}$ are i.i.d. Gaussian variables with mean zero and variance σ^2, we have that each S_m is a Gaussian variable with mean zero and variance given by $\operatorname{Var}(S_m):=\sum_{\ell=-n}^{n}\left(\frac{g_M(\ell)}{M}\right)^2\sigma^2$. The main idea next is first to compute an upper bound (denoted by $\overline{\Pi}$) on the variance $\operatorname{Var}(S_m)$ and then apply the Gaussian upper deviation inequality [268, Eq. (7.8)]

$$\mathbb{P}\left[|S_m|\geq t\sqrt{\overline{\Pi}}\right]\leq e^{-t^2/2} \tag{A.22}$$

to get a high-probability upper bound on $|S_m|$. To evaluate $\overline{\Pi}$, it is instructive to first note

$$g_M(\ell) = \frac{1}{M} \sum_{k=\max(\ell-M,-M)}^{\min(\ell+M,M)} \left(1 - \left|\frac{k}{M}\right|\right)\left(1 - \left|\frac{\ell-k}{M}\right|\right),$$

with $\ell = -2M, \ldots, 2M$, which is the convolution of two triangle functions:

$$g_M(\ell) = \frac{1}{M} \mathrm{Tri}_M(\ell) * \mathrm{Tri}_M(\ell),\ \ell = -2M, \ldots, 2M. \tag{A.23}$$

Here the triangle function is defined by $\mathrm{Tri}_M(\ell) := 1 - \frac{|\ell|}{M},\ \ell = -M, \ldots, M$ and $*$ represents the convolution operator. Apparently $\mathrm{Var}(S_m)$ is the squared ℓ_2 norm of the vector $\mathbf{g}_M := [g_M(-2M), \ldots, g_M(2M)]^T$ scaled by σ^2/M^2. Since by Eq. (A.23), $\mathbf{g}_M$ is the convolution of two (the same) triangular vectors $\mathbf{h}_M := [\mathrm{Tri}_M(-M), \ldots, \mathrm{Tri}_M(M)]$ and then scaled by $1/M$, we obtain an upper bound on $\mathrm{Var}(S_m)$ by applying Young's inequality $\|\mathbf{f} * \mathbf{g}\|_r \le \|\mathbf{f}\|_p \|\mathbf{g}\|_q$ where $r^{-1} = p^{-1} + q^{-1} - 1$ and setting $r = 2, p = 2, q = 1$:

$$\mathrm{Var}(S_m) = \sum_{\ell=-n}^{n} \left(\frac{g_M(\ell)}{M}\right)^2 \sigma^2 = \frac{\sigma^2}{M^2}\|\mathbf{g}_M\|_2^2 = \frac{\sigma^2}{M^4}\|\mathbf{h}_M * \mathbf{h}_M\|_2^2 \le \frac{\sigma^2}{M^4}\|\mathbf{h}_M\|_1^2 \|\mathbf{h}_M\|_2^2. \tag{A.24}$$

Therefore, to bound $\mathrm{Var}(S_m)$, it remains to bound $\|\mathbf{h}_M\|_1^2$ and $\|\mathbf{h}_M\|_2^2$:

$$\|\mathbf{h}_M\|_1^2 = \left(\sum_{\ell=-M}^{M}\left(1 - \frac{|\ell|}{M}\right)\right)^2 = M^2 \qquad \text{and} \qquad \|\mathbf{h}_M\|_2^2 = \sum_{\ell=-M}^{M}\left(1 - \frac{|\ell|}{M}\right)^2 = \frac{2M}{3} + \frac{1}{3M}. \tag{A.25}$$

Then plug (A.25) into (A.24), we obtain an upper bound on $\mathrm{Var}(S_m)$ as

$$\mathrm{Var}(S_m) \le \frac{\sigma^2}{M^4} M^2 \left(\frac{2M}{3} + \frac{1}{3M}\right) = \sigma^2 \left(\frac{2}{3M} + \frac{1}{M^3}\right) \overset{①}{=} \sigma^2 \left(\frac{4}{3n} + \frac{8}{n^3}\right) \overset{②}{\le} 1.334\sigma^2/n, \quad \text{for } n \ge 130, \tag{A.26}$$

where ① follows from $n = 2M$ and ② follows since $8/n^2$ is a decreasing sequence of n implying the maximal happens at $n = 130$. Thus we can choose $\overline{\Pi} = 1.334\sigma^2/n$. Plugging $\overline{\Pi} = 1.334\sigma^2/n$ into the Gaussian tail bound (A.22), we get

$$\mathbb{P}\left[|S_m| \ge t\sqrt{1.334}\sigma/\sqrt{n}\right] \le e^{-t^2/2} \tag{A.27}$$

for all $m = 0, \ldots, N-1$.

Applying the union bound yields

$$\mathbb{P}\left[\sup_{f\in\mathbb{T}} \left|\mathbf{a}(f)^H \mathbf{w}\right| \ge 2t\sqrt{1.334}\sigma/\sqrt{n}\right] \le \mathbb{P}\left[\max_{m=0\ldots N-1} |S_m| \ge t\sqrt{1.334}\sigma/\sqrt{n}\right] \le N e^{-t^2/2}, \tag{A.28}$$

where the first inequality follows from (A.21). Setting $t = \sqrt{8\log n}$ in the above gives

$$\mathbb{P}\left[\sup_{f\in\mathbb{T}}\left|\mathbf{a}(f)^H\mathbf{w}\right| \geq \underbrace{4\sqrt{2}\sqrt{1.334}}_{\leq 6.534}\sqrt{\log n/n}\sigma\right] \leq \frac{8\pi(2n+1)}{n^4} \leq \frac{1}{n^2}, \tag{A.29}$$

where the last inequality holds for $n \geq 130$. Therefore, we obtain that

$$\mathbb{P}\left[\sup_{f\in\mathbb{T}}\left|\mathbf{a}(f)^H\mathbf{w}\right| \leq 6.534\sqrt{\log n/n}\sigma\right] \geq 1-\frac{1}{n^2}, \quad \text{for } n \geq 130. \tag{A.30}$$

To bound $\sup_{f\in\mathbb{T}}\left|\mathbf{a}'(f)^H\mathbf{Z}\mathbf{w}\right|$ and $\sup_{f\in\mathbb{T}}\left|\mathbf{a}''(f)^H\mathbf{Z}\mathbf{w}\right|$, a natural approach is to exploit the relations between $\mathbf{a}(\mathbf{f})$ and its derivatives $\mathbf{a}'(\mathbf{f})$, $\mathbf{a}''(\mathbf{f})$:

$$\mathbf{a}'(f) = (i2\pi\,\mathrm{diag}(\mathbf{n})))\mathbf{a}(f) \text{ and } \mathbf{a}''(f) = (i2\pi\,\mathrm{diag}(\mathbf{n}))^2\mathbf{a}(f).$$

Similarly, define S'_m and S''_m as the mth equispaced sample of $\mathbf{a}'(f)^H\mathbf{Z}\mathbf{w}$ and $\mathbf{a}''(f)^H\mathbf{Z}\mathbf{w}$, respectively:

$$\begin{aligned} S'_m &= \mathbf{a}'(m/N)^H\mathbf{Z}\mathbf{w} = \mathbf{a}(m/N)^H(-i2\pi\,\mathrm{diag}(\mathbf{n}))\mathbf{Z}\mathbf{w}, \\ S''_m &= \mathbf{a}''(m/N)^H\mathbf{Z}\mathbf{w} = \mathbf{a}(m/N)^H(-i2\pi\,\mathrm{diag}(\mathbf{n}))^2\mathbf{Z}\mathbf{w}. \end{aligned}$$

Hence $S'_m \sim \mathcal{N}(0, \mathrm{Var}(S'_m))$ and $S''_m \sim \mathcal{N}(0, \mathrm{Var}(S''_m))$ with

$$\begin{aligned} \mathrm{Var}(S'_m) &= \sum_{\ell=-n}^{n}\left(2\pi\ell g_M(\ell)/M\right)^2\sigma^2) \leq (2\pi n)^2\left(\sum_{\ell=-n}^{n}\left(g_M(\ell)/M\right)^2\sigma^2\right) \overset{①}{\leq} (2\pi n)^2 1.334\sigma^2/n, \\ \mathrm{Var}(S''_m) &= \sum_{\ell=-n}^{n}\left((2\pi\ell)^2 g_M(\ell)/M\right)^2\sigma^2) \leq (2\pi n)^4\left(\sum_{\ell=-n}^{n}\left(g_M(\ell)/M\right)^2\sigma^2\right) \overset{②}{\leq} (2\pi n)^4 1.334\sigma^2/n, \end{aligned}$$

where ① and ② follow from (A.26). Applying the Gaussian deviation inequality to S'_m, S''_m yields

$$\mathbb{P}\left[|S'_m| \geq t2\pi\sqrt{1.334}\sqrt{n}\sigma\right] \leq 2e^{-t^2/2} \qquad \text{and} \qquad \mathbb{P}\left[|S''_m| \geq t4\pi^2\sqrt{1.334}n\sqrt{n}\sigma\right] \leq 2e^{-t^2/2}.$$

Then applying the same arguments as (A.28), we get for $n \geq 130$,

$$\begin{aligned} &\mathbb{P}\left[\sup_{f\in\mathbb{T}}\left|\mathbf{a}'(f)^H\mathbf{w}\right| \leq \underbrace{8\sqrt{2}\pi\sqrt{1.334}}_{\leq 41.052}\sqrt{n\log n}\sigma\right] \geq 1-\frac{1}{n^2}, \\ &\mathbb{P}\left[\sup_{f\in\mathbb{T}}\left|\mathbf{a}''(f)^H\mathbf{w}\right| \leq \underbrace{16\sqrt{2}\pi^2\sqrt{1.334}}_{\leq 257.94}n\sqrt{n\log n}\sigma\right] \geq 1-\frac{1}{n^2}. \end{aligned} \tag{A.31}$$

Finally, we invoke that

$$\mathbf{A}(\mathbf{f}) = [\mathbf{a}(f_1), \ldots, \mathbf{a}(f_k)], \qquad \mathbf{A}'(\mathbf{f}) = [\mathbf{a}'(f_1), \ldots, \mathbf{a}'(f_k)], \qquad \mathbf{A}''(\mathbf{f}) = [\mathbf{a}''(f_1), \ldots, \mathbf{a}''(f_k)],$$

and recognize that $\sup_{f\in\mathbb{T}} \left|\mathbf{a}^{(\ell)}(f)^H\mathbf{w}\right|$ is an upper bound on $\|\mathbf{A}^{(\ell)}(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty$ to get

$$\|\mathbf{A}^{(\ell)}(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty = \max_{f\in\{f_j\}} |\mathbf{a}^{(\ell)}(f)^H\mathbf{Z}\mathbf{w}| \le \sup_{f\in\mathbb{T}} |\mathbf{a}^{(\ell)}(f)^H\mathbf{Z}\mathbf{w}|.$$

Together with (A.30), (A.31) and the definition $\gamma_0 = \sigma\sqrt{\frac{\log n}{n}}$, we obtain that the following inequalities hold for $n \ge 130$ with probability at least $1 - \frac{1}{n^2}$:

$$\begin{aligned}
\|\mathbf{A}(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty &\le \sup_{f\in\mathbb{T}} \left|\mathbf{a}(f)^H\mathbf{w}\right| \le 6.534\gamma_0, \\
\|\mathbf{A}'(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty &\le \sup_{f\in\mathbb{T}} \left|\mathbf{a}'(f)^H\mathbf{w}\right| \le 41.052n\gamma_0, \\
\|\mathbf{A}''(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty &\le \sup_{f\in\mathbb{T}} \left|\mathbf{a}''(f)^H\mathbf{w}\right| \le 257.94n^2\gamma_0.
\end{aligned} \tag{A.32}$$

As a consequence, we claim that the following inequalities hold for $n \ge 130$ with probability at least $1 - \frac{1}{n^2}$:

$$\begin{aligned}
\|\operatorname{diag}(1./|\mathbf{c}^\star|)\mathbf{A}'(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty/\sqrt{\tau} &\overset{①}{\le} \|\operatorname{diag}(1./|\mathbf{c}^\star|)\|_{\infty,\infty}\|\mathbf{A}(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty/\sqrt{\tau} \\
&\le \frac{1}{\sqrt{3.289n^2}}\frac{1}{c^\star_{\min}}(41.052n\gamma_0) \le 22.64\gamma,
\end{aligned} \tag{A.33}$$

$$\begin{aligned}
\|\operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|^2)\mathbf{A}''(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty/\tau &\overset{②}{\le} \|\operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|)\|_{\infty,\infty}\|\operatorname{diag}(1./|\mathbf{c}^\star|)\|_{\infty,\infty}\|\mathbf{A}''(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty/\tau \\
&\le \frac{1}{3.289n^2}(1+X^\star\gamma)\frac{1}{c^\star_{\min}}(257.94n^2\gamma_0) \le 78.43(1+X^\star\gamma)\gamma,
\end{aligned} \tag{A.34}$$

where ① follows from that $\|\mathbf{A}\mathbf{x}\|_\infty \le \|\mathbf{A}\|_{\infty,\infty}\|\mathbf{x}\|_\infty$ by the definition of the $\ell_{\infty,\infty}$ norm and the fact $\tau \ge 3.289n^2$ for $n \ge 130$ by (A.2). ② follows from the sub-multiplicative property of the $\ell_{\infty,\infty}$ norm that

$$\|\mathbf{A}\mathbf{B}\mathbf{x}\|_\infty \le \|\mathbf{A}\|_{\infty,\infty}\|\mathbf{B}\|_{\infty,\infty}\|\mathbf{x}\|_\infty$$

and $\|\operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|)\|_{\infty,\infty} = \max_\ell |c_\ell|/|c^\star_\ell| \le (1+X^\star\gamma)$ which follows from the assumption $\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\|_\infty \le X^\star\gamma_0/\sqrt{2}$ and the derived results (A.10).

A.3 Gradient and Hessian for the Nonconvex Program (2.15)

Recall that the objective function G of the program (2.15) is

$$G(\mathbf{f},\mathbf{c}) = \frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{y}\|_Z^2 + \lambda\|\mathbf{c}\|_1.$$

We denote $\mathbf{c} = \mathbf{u} + i\mathbf{v}$ for $\mathbf{u} \in \mathbb{R}^k$ and $\mathbf{v} \in \mathbb{R}^k$.

A.3.1 Gradient

Let the operators $\mathbb{R}\{\cdot\}$ and $\mathbb{I}\{\cdot\}$ take respectively the real and imaginary parts of a complex number or vector. The gradient of $G(\mathbf{f},\mathbf{c})$ with respect to $\boldsymbol{\theta} := (\mathbf{f},\mathbf{u},\mathbf{v}) \in \mathbb{R}^{3k}$ is defined by

$$\nabla G(\boldsymbol{\theta}) = \begin{bmatrix} \partial G/\partial \mathbf{f} \\ \partial G/\partial \mathbf{u} \\ \partial G/\partial \mathbf{v} \end{bmatrix} \overset{①}{=} \begin{bmatrix} \partial G/\partial \mathbf{f} \\ 2\mathbb{R}\{\partial G/\partial \bar{\mathbf{c}}\} \\ 2\mathbb{I}\{\partial G/\partial \bar{\mathbf{c}}\} \end{bmatrix} \overset{②}{=} \begin{bmatrix} \mathbb{R}\left\{(\mathbf{A}'(\mathbf{f})\,\mathrm{diag}(\mathbf{c}))^H \mathbf{Z}(\mathbf{A}(\mathbf{f})\mathbf{c}-\mathbf{y})\right\} \\ \mathbb{R}\left\{\mathbf{A}(\mathbf{f})^H \mathbf{Z}(\mathbf{A}(\mathbf{f})\mathbf{c}-\mathbf{y}) + \lambda \mathbf{c}./|\mathbf{c}|\right\} \\ \mathbb{I}\left\{\mathbf{A}(\mathbf{f})^H \mathbf{Z}(\mathbf{A}(\mathbf{f})\mathbf{c}-\mathbf{y}) + \lambda \mathbf{c}./|\mathbf{c}|\right\} \end{bmatrix}$$

$$\overset{③}{=} \begin{bmatrix} \mathbb{R}\left\{\mathrm{diag}(\mathbf{c})^H(-\mathbf{D}_1(\mathbf{f})\mathbf{c} + \mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{A}'(\mathbf{f})^H \mathbf{Z}\mathbf{w})\right\} \\ \mathbb{R}\left\{\mathbf{D}_0(\mathbf{f})\mathbf{c} - \mathbf{D}_0(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{A}(\mathbf{f})^H \mathbf{Z}\mathbf{w} + \lambda \mathbf{c}./|\mathbf{c}|\right\} \\ \mathbb{I}\left\{\mathbf{D}_0(\mathbf{f})\mathbf{c} - \mathbf{D}_0(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{A}(\mathbf{f})^H \mathbf{Z}\mathbf{w} + \lambda \mathbf{c}./|\mathbf{c}|\right\} \end{bmatrix}, \quad \text{(A.35)}$$

where ① holds for $G \in \mathbb{R}$. ② follows from $\mathrm{diag}(\mathrm{d}\mathbf{f})\mathbf{c} = \mathrm{diag}(\mathbf{c})\mathrm{d}\mathbf{f}$ and $\mathrm{d}|c| = \frac{\bar{c}\mathrm{d}c + c\mathrm{d}\bar{c}}{2|c|}$. ③ follows from the kernel matrix factorization formulas (A.6)- (A.7) and by taking into account that $\mathbf{y} = \mathbf{x}^\star + \mathbf{w} = \mathbf{A}(\mathbf{f}^\star)\mathbf{c}^\star + \mathbf{w}$.

A.3.2 Hessian

The symmetric Hessian matrix $\nabla^2 G(\boldsymbol{\theta})$ is given by

$$\nabla^2 G(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2 G}{\partial \mathbf{f}^T \partial \mathbf{f}} & \frac{\partial^2 G}{\partial \mathbf{f}^T \partial \mathbf{u}} & \frac{\partial^2 G}{\partial \mathbf{f}^T \partial \mathbf{v}} \\ \frac{\partial^2 G}{\partial \mathbf{u}^T \partial \mathbf{f}} & \frac{\partial^2 G}{\partial \mathbf{u}^T \partial \mathbf{u}} & \frac{\partial^2 G}{\partial \mathbf{u}^T \partial \mathbf{v}} \\ \frac{\partial^2 G}{\partial \mathbf{v}^T \partial \mathbf{f}} & \frac{\partial^2 G}{\partial \mathbf{v}^T \partial \mathbf{u}} & \frac{\partial^2 G}{\partial \mathbf{v}^T \partial \mathbf{v}} \end{bmatrix} := \begin{bmatrix} \mathbf{H}_{\mathbf{ff}} & \mathbf{H}_{\mathbf{fu}} & \mathbf{H}_{\mathbf{fv}} \\ \mathbf{H}_{\mathbf{uf}} & \mathbf{H}_{\mathbf{uu}} & \mathbf{H}_{\mathbf{uv}} \\ \mathbf{H}_{\mathbf{vf}} & \mathbf{H}_{\mathbf{vu}} & \mathbf{H}_{\mathbf{vv}} \end{bmatrix}$$

with

$$\begin{bmatrix} \mathbf{H}_{\mathbf{ff}} \\ \mathbf{H}_{\mathbf{fu}} \\ \mathbf{H}_{\mathbf{fv}} \\ \mathbf{H}_{\mathbf{uu}} \\ \mathbf{H}_{\mathbf{vv}} \\ \mathbf{H}_{\mathbf{uv}} \end{bmatrix} \overset{①}{=} \begin{bmatrix} \mathbb{R}\{(\mathbf{A}'(\mathbf{f})\boldsymbol{\Lambda})^H \mathbf{Z}\mathbf{A}'(\mathbf{f})\boldsymbol{\Lambda} + \mathrm{diag}((\mathbf{A}''(\mathbf{f})\boldsymbol{\Lambda})^H \mathbf{Z}(\mathbf{A}(\mathbf{f})\mathbf{c}-\mathbf{y}))\} \\ \mathbb{R}\{(\mathbf{A}'(\mathbf{f})\boldsymbol{\Lambda})^H \mathbf{Z}\mathbf{A}(\mathbf{f}) + \mathrm{diag}(\mathbf{A}'(\mathbf{f})^H \mathbf{Z}(\mathbf{A}(\mathbf{f})\mathbf{c}-\mathbf{y}))\} \\ \mathbb{I}\{-(\mathbf{A}'(\mathbf{f})\boldsymbol{\Lambda})^H \mathbf{Z}\mathbf{A}(\mathbf{f}) + \mathrm{diag}(\mathbf{A}'(\mathbf{f})^H \mathbf{Z}(\mathbf{A}(\mathbf{f})\mathbf{c}-\mathbf{y}))\} \\ \mathbf{A}(\mathbf{f})^H \mathbf{Z}\mathbf{A}(\mathbf{f}) + \lambda\,\mathrm{diag}(\mathbf{v}^2 \odot |\mathbf{c}|.^{-3}) \\ \mathbf{A}(\mathbf{f})^H \mathbf{Z}\mathbf{A}(\mathbf{f}) + \lambda\,\mathrm{diag}(\mathbf{u}^2 \odot |\mathbf{c}|.^{-3}) \\ -\lambda\,\mathrm{diag}(\mathbf{u} \odot \mathbf{v} \odot |\mathbf{c}|.^{-3}) \end{bmatrix}$$

$$\overset{②}{=} \begin{bmatrix} \mathbb{R}\{-\boldsymbol{\Lambda}^H \mathbf{D}_2(\mathbf{f})\boldsymbol{\Lambda} - \mathrm{diag}(\boldsymbol{\Lambda}^H \mathbf{A}''(\mathbf{f})^H \mathbf{Z}\mathbf{w}) - \mathrm{diag}(\boldsymbol{\Lambda}^H(\mathbf{D}_2(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_2(\mathbf{f})\mathbf{c}))\} \\ \mathbb{R}\{-\boldsymbol{\Lambda}^H \mathbf{D}_1(\mathbf{f}) - \mathrm{diag}(\mathbf{A}'(\mathbf{f})^H \mathbf{Z}\mathbf{w}) + \mathrm{diag}(\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star) - \mathrm{diag}(\mathbf{D}_1(\mathbf{f})\mathbf{c})\} \\ \mathbb{I}\{\boldsymbol{\Lambda}^H \mathbf{D}_1(\mathbf{f}) - \mathrm{diag}(\mathbf{A}'(\mathbf{f})^H \mathbf{Z}\mathbf{w}) + \mathrm{diag}(\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star) - \mathrm{diag}(\mathbf{D}_1(\mathbf{f})\mathbf{c})\} \\ \mathbf{D}_0(\mathbf{f}) + \lambda\,\mathrm{diag}(\mathbf{v}^2 \odot |\mathbf{c}|.^{-3}) \\ \mathbf{D}_0(\mathbf{f}) + \lambda\,\mathrm{diag}(\mathbf{u}^2 \odot |\mathbf{c}|.^{-3}) \\ -\lambda\,\mathrm{diag}(\mathbf{u} \odot \mathbf{v} \odot |\mathbf{c}|.^{-3}) \end{bmatrix}, \quad \text{(A.36)}$$

where we denoted $\boldsymbol{\Lambda} := \mathrm{diag}(\mathbf{c})$ to simplify notation. ① follows from direct computation and ② follows from the matrix decomposition formulas (A.6)- (A.7) and by taking into account that $\mathbf{y} = \mathbf{x}^\star + \mathbf{w} = \mathbf{A}(\mathbf{f}^\star)\mathbf{c}^\star + \mathbf{w}$.

Remarkably, if we replace the noisy signal $\mathbf{y}$ in the objective function of the nonconvex program (2.15) with the noise-free signal $\mathbf{x}^\star$ to get

$$G^\lambda(\mathbf{f},\mathbf{c}) = \frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{x}^\star\|_Z^2 + \lambda\|\mathbf{c}\|_1,$$

then its gradient and Hessian matrix can be obtained from those of $G(\mathbf{f},\mathbf{c})$ by setting the noise $\mathbf{w}$ to zero.

A.4 Proof of Lemma 2.4.1

Lemma A.4.1 (Lemma 2.4.1)**.** *Let the first fixed point map be the weighted gradient map of the nonconvex program* (2.15) *with the noisy signal* $\mathbf{y}$ *replaced by the noise-free signal* $\mathbf{x}^\star$*:*

$$\Theta^\lambda(\boldsymbol{\theta}) := \boldsymbol{\theta} - \mathbf{W}^\star \nabla \left(\frac{1}{2} \|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{x}^\star\|_{\mathbf{Z}}^2 + \lambda \|\mathbf{c}\|_1 \right), \tag{2.19}$$

where the gradient ∇ *is taken with respect to the parameter* $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{u}, \mathbf{v})$*. Let the regularization parameter* λ *vary in* $[0, 0.646X^\star\gamma_0]$*. Define a neighborhood* $\mathcal{N}^\star := \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_{\hat{\infty}} \le X^\star\gamma_0/\sqrt{2}\right\}$*. Suppose that the separation condition* (2.9) *and the SNR condition* (2.10) *hold. Then the map* Θ^λ *has a unique fixed point* $\boldsymbol{\theta}^\lambda \in \mathcal{N}^\star$ *satisfying* $\Theta^\lambda(\boldsymbol{\theta}^\lambda) = \boldsymbol{\theta}^\lambda$*. Furthermore, according to the implicit function theorem,* $\boldsymbol{\theta}^\lambda$ *is a continuously differentiable function of* λ *whose derivative is given by*

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\boldsymbol{\theta}^\lambda = -(\nabla^2 G^\lambda(\boldsymbol{\theta}^\lambda))^{-1} \frac{\partial}{\partial \lambda} \nabla G^\lambda(\boldsymbol{\theta}^\lambda). \tag{2.20}$$

Finally, when λ *turns to zero, the fixed point* $\boldsymbol{\theta}^\lambda$ *converges to* $\boldsymbol{\theta}^\star$*, i.e.,* $\lim_{\lambda\to 0} \boldsymbol{\theta}^\lambda = \boldsymbol{\theta}^\star$*, and therefore* $\lim_{\lambda\to 0} \mathbf{x}^\lambda = \mathbf{x}^\star$*.*

Proof. The underlying fixed point map is

$$\Theta^\lambda(\boldsymbol{\theta}) = \boldsymbol{\theta} - \mathbf{W}^\star \nabla G^\lambda(\boldsymbol{\theta}),$$

where G^λ is defined as the objective function of the nonconvex program (2.15) with the noisy signal $\mathbf{y}$ replaced by the noise-free signal $\mathbf{x}^\star$:

$$G^\lambda(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{x}^\star\|_{\mathbf{Z}}^2 + \lambda\|\mathbf{c}\|_1.$$

By Theorem 2.4.1, to show the existence and uniqueness of a point $\boldsymbol{\theta}^\lambda \in \mathcal{N}^\star$ such that $\Theta^\lambda(\boldsymbol{\theta}^\lambda) = \boldsymbol{\theta}^\lambda$, the key is to show that Θ^λ satisfies the non-escaping condition and the contraction condition:

(i) $\Theta^\lambda(\mathcal{N}^\star) \subset \mathcal{N}^\star$;
(ii) There exists $\rho \in (0,1)$ such that $\|\Theta^\lambda(\mathbf{v}) - \Theta^\lambda(\mathbf{w})\|_{\hat{\infty}} \le \rho\|\mathbf{v} - \mathbf{w}\|_{\hat{\infty}}$ for any $\mathbf{v}, \mathbf{w} \in \mathcal{N}^\star$.

A.4.1 Showing the Contraction Property

For $\mathbf{v}, \mathbf{w} \in \mathcal{N}^\star$, we have

$$\|\Theta^\lambda(\mathbf{v}) - \Theta^\lambda(\mathbf{w})\|_{\hat{\infty}} \overset{①}{=} \left\| \int_0^1 [\nabla\Theta^\lambda(t\mathbf{v} + (1-t)\mathbf{w})](\mathbf{v} - \mathbf{w})\mathrm{d}t \right\|_{\hat{\infty}} \overset{②}{\le} \underset{\boldsymbol{\theta}\in\mathcal{N}^\star}{\text{maximize}} \|\nabla\Theta^\lambda(\boldsymbol{\theta})\|_{\hat{\infty},\hat{\infty}} \|\mathbf{v} - \mathbf{w}\|_{\hat{\infty}},$$

where ① follows from the integral form of the mean value theorem for vector-valued functions (see [148, Eq. (A.57)]); ② follows from the sub-multiplicative property of $\|\cdot\|_{\hat{\infty},\hat{\infty}}$ and the fact that $t\mathbf{v} + (1-t)\mathbf{w} \in \mathcal{N}^\star$ for $t \in [0,1]$. Thus,

it suffices to show

$$\underset{\boldsymbol{\theta}\in\mathcal{N}^\star}{\text{maximize}}\,\|\nabla\Theta^\lambda(\boldsymbol{\theta})\|_{\hat{\infty},\hat{\infty}}<1,$$

where the matrix $\ell_{\hat{\infty},\hat{\infty}}$ norm is defined by (following from the definition of the $\ell_{\hat{\infty}}$ norm)

$$\|\mathbf{A}\|_{\hat{\infty},\hat{\infty}}=\left\|\begin{bmatrix}\mathbf{A}_{11}&\mathbf{A}_{12}&\mathbf{A}_{13}\\\mathbf{A}_{21}&\mathbf{A}_{22}&\mathbf{A}_{23}\\\mathbf{A}_{31}&\mathbf{A}_{32}&\mathbf{A}_{33}\end{bmatrix}\right\|_{\hat{\infty},\hat{\infty}}:=\left\|\begin{bmatrix}\mathbf{S}\mathbf{A}_{11}\mathbf{S}^{-1}&\mathbf{S}\mathbf{A}_{12}&\mathbf{S}\mathbf{A}_{13}\\\mathbf{A}_{21}\mathbf{S}^{-1}&\mathbf{A}_{22}&\mathbf{A}_{23}\\\mathbf{A}_{31}\mathbf{S}^{-1}&\mathbf{A}_{32}&\mathbf{A}_{33}\end{bmatrix}\right\|_{\infty,\infty},$$

with $\mathbf{S}=\sqrt{\tau}\,\mathrm{diag}(|\mathbf{c}^\star|)$. Together with

$$\mathbf{W}^\star=\begin{bmatrix}\mathbf{S}^{-2}&&\\&\mathbf{I}_k&\\&&\mathbf{I}_k\end{bmatrix},$$

we therefore obtain that

$$\|\mathbf{W}^\star\mathbf{A}\|_{\hat{\infty},\hat{\infty}}=\left\|\begin{bmatrix}\mathbf{S}^{-1}\mathbf{A}_{11}\mathbf{S}^{-1}&\mathbf{S}^{-1}\mathbf{A}_{12}&\mathbf{S}^{-1}\mathbf{A}_{13}\\\mathbf{A}_{21}\mathbf{S}^{-1}&\mathbf{A}_{22}&\mathbf{A}_{23}\\\mathbf{A}_{31}\mathbf{S}^{-1}&\mathbf{A}_{32}&\mathbf{A}_{33}\end{bmatrix}\right\|_{\infty,\infty}=\|\mathbf{W}^{\star\frac{1}{2}}\mathbf{A}\mathbf{W}^{\star\frac{1}{2}}\|_{\infty,\infty}:=\|\Upsilon(\mathbf{A})\|_{\infty,\infty},\tag{A.37}$$

where the linear operator $\Upsilon(\cdot):=\mathbf{W}^{\star\frac{1}{2}}(\cdot)\mathbf{W}^{\star\frac{1}{2}}$. The Jacobian of the fixed point map Θ^λ is given by

$$\nabla\Theta^\lambda(\boldsymbol{\theta})=\mathbf{I}-\mathbf{W}^\star\nabla^2G^\lambda(\boldsymbol{\theta}),\tag{A.38}$$

where the symmetric Hessian matrix $\nabla^2G^\lambda(\boldsymbol{\theta})$ can be obtained from $\nabla^2G(\boldsymbol{\theta})$ by setting the noise $\mathbf{w}$ to zero:

$$\nabla^2G^\lambda(\boldsymbol{\theta})=\begin{bmatrix}\mathbf{H}_{\mathbf{ff}}&\mathbf{H}_{\mathbf{fu}}&\mathbf{H}_{\mathbf{fv}}\\\mathbf{H}_{\mathbf{uf}}&\mathbf{H}_{\mathbf{uu}}&\mathbf{H}_{\mathbf{uv}}\\\mathbf{H}_{\mathbf{vf}}&\mathbf{H}_{\mathbf{vu}}&\mathbf{H}_{\mathbf{vv}}\end{bmatrix}.$$

Due to the symmetric structure of the Hessian matrix, it suffices to know the expressions for the following block matrices (see Eq. (A.36)):

$$\begin{aligned}
&\mathbf{H}_{\mathbf{ff}}=\mathbb{R}\{-\boldsymbol{\Lambda}^H\mathbf{D}_2(\mathbf{f})\boldsymbol{\Lambda}-\mathrm{diag}(\boldsymbol{\Lambda}^H(\mathbf{D}_2(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star-\mathbf{D}_2(\mathbf{f})\mathbf{c}))\};\quad &&\mathbf{H}_{\mathbf{uu}}=\mathbf{D}_0(\mathbf{f})+\lambda\,\mathrm{diag}(\mathbf{v}\odot\mathbf{v}./|\mathbf{c}|^3);\\
&\mathbf{H}_{\mathbf{fu}}=\mathbb{R}\{-\boldsymbol{\Lambda}^H\mathbf{D}_1(\mathbf{f})+\mathrm{diag}(\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star)-\mathrm{diag}(\mathbf{D}_1(\mathbf{f})\mathbf{c})\};\quad &&\mathbf{H}_{\mathbf{vv}}=\mathbf{D}_0(\mathbf{f})+\lambda\,\mathrm{diag}(\mathbf{u}\odot\mathbf{u}./|\mathbf{c}|^3);\\
&\mathbf{H}_{\mathbf{fv}}=\mathbb{I}\{\boldsymbol{\Lambda}^H\mathbf{D}_1(\mathbf{f})+\mathrm{diag}(\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star)-\mathrm{diag}(\mathbf{D}_1(\mathbf{f})\mathbf{c})\};\quad &&\mathbf{H}_{\mathbf{uv}}=-\lambda\,\mathrm{diag}(\mathbf{u}\odot\mathbf{v}./|\mathbf{c}|^3),
\end{aligned}$$

where $\boldsymbol{\Lambda}=\mathrm{diag}(\mathbf{c})$.

Next we compute the weighed $\ell_{\hat{\infty},\hat{\infty}}$ norm of the Jacobian of the fixed point map Θ^λ:

$$\|\nabla\Theta^\lambda(\boldsymbol{\theta})\|_{\hat{\infty},\hat{\infty}}\overset{①}{=}\|\mathbf{W}^\star\nabla^2G^\lambda(\boldsymbol{\theta})-\mathbf{I}\|_{\hat{\infty},\hat{\infty}}\overset{②}{=}\|\Upsilon(\nabla^2G^\lambda(\boldsymbol{\theta})-\mathbf{W}^{\star-1})\|_{\infty,\infty}\overset{③}{=}\|\Upsilon(\nabla^2G^\lambda(\boldsymbol{\theta}))-\mathbf{I}\|_{\infty,\infty},$$

where ① follows from (A.38), ② follows from (A.37) by noting that $\mathbf{W}^\star\nabla^2 G^\lambda(\boldsymbol{\theta}) - \mathbf{I} = \mathbf{W}^\star(\nabla^2 G^\lambda(\boldsymbol{\theta}) - \mathbf{W}^{\star -1})$ and ③ from the linearity of $\Upsilon(\cdot)$ and $\Upsilon(\mathbf{W}^{\star -1}) = \mathbf{W}^{\star \frac{1}{2}}\mathbf{W}^{\star -1}\mathbf{W}^{\star \frac{1}{2}} = \mathbf{I}$. Direct computation gives

$$\Upsilon(\nabla^2 G^\lambda(\boldsymbol{\theta}))-\mathbf{I}=\begin{bmatrix} \frac{-1}{\tau}\mathbb{R}\{\boldsymbol{\Phi}^H\mathbf{D}_2(\mathbf{f})\boldsymbol{\Phi}\}-\mathbf{I} & \frac{-1}{\sqrt{\tau}}\mathbb{R}\{\boldsymbol{\Phi}\}\mathbf{D}_1(\mathbf{f}) & \frac{-1}{\sqrt{\tau}}\mathbb{I}\{\boldsymbol{\Phi}\}\mathbf{D}_1(\mathbf{f}) \\ \frac{1}{\sqrt{\tau}}\mathbf{D}_1(\mathbf{f})\mathbb{R}\{\boldsymbol{\Phi}\} & \mathbf{D}_0(\mathbf{f})-\mathbf{I} & \\ \frac{1}{\sqrt{\tau}}\mathbf{D}_1(\mathbf{f})\mathbb{I}\{\boldsymbol{\Phi}\} & & \mathbf{D}_0(\mathbf{f})-\mathbf{I} \end{bmatrix} + \begin{bmatrix} \mathrm{diag}(\mathbf{d_{ff}}) & \mathrm{diag}(\mathbf{d_{fu}}) & \mathrm{diag}(\mathbf{d_{fv}}) \\ \mathrm{diag}(\mathbf{d_{fu}}) & \mathrm{diag}(\mathbf{d_{uu}}) & \mathrm{diag}(\mathbf{d_{uv}}) \\ \mathrm{diag}(\mathbf{d_{fv}}) & \mathrm{diag}(\mathbf{d_{uv}}) & \mathrm{diag}(\mathbf{d_{vv}}) \end{bmatrix}$$

where $\boldsymbol{\Phi} := \mathrm{diag}(\mathbf{c}./|\mathbf{c}^\star|)$ and

$$\begin{aligned} \mathbf{d_{ff}} &= -\mathbb{R}\{\mathrm{diag}(\mathbf{c}./|\mathbf{c}^\star|^2)^H[\mathbf{D}_2(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_2(\mathbf{f})\mathbf{c}]/\tau\}; \quad & \mathbf{d_{uu}} &= \lambda\,\mathrm{diag}(\mathbf{u}\odot\mathbf{u}./\left|\mathbf{c}\right|^3); \\ \mathbf{d_{fu}} &= \mathbb{R}\{\mathrm{diag}(1./|\mathbf{c}^\star|)[\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f})\mathbf{c}]/\sqrt{\tau}\}; & \mathbf{d_{uv}} &= \lambda\,\mathrm{diag}(\mathbf{u}\odot\mathbf{v}./\left|\mathbf{c}\right|^3); \\ \mathbf{d_{fv}} &= \mathbb{I}\{\mathrm{diag}(1./|\mathbf{c}^\star|)[\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f})\mathbf{c}]/\sqrt{\tau}\}; & \mathbf{d_{vv}} &= \lambda\,\mathrm{diag}(\mathbf{v}\odot\mathbf{v}./\left|\mathbf{c}\right|^3). \end{aligned}$$

Clearly,

$$\|\Upsilon(\nabla^2 G^\lambda(\boldsymbol{\theta})) - \mathbf{I}\|_{\infty,\infty} = \max\left\{\Pi_1^\lambda, \Pi_2^\lambda, \Pi_3^\lambda\right\}$$

with $\Pi_1^\lambda, \Pi_2^\lambda, \Pi_3^\lambda$ being the first, second and third absolute row sums of $\Upsilon(\nabla^2 G^\lambda(\boldsymbol{\theta})) - \mathbf{I}$, respectively.

Bounding Π_1^λ.

$$\begin{aligned} \Pi_1^\lambda \leq& \left\|-\mathbb{R}\{\mathrm{diag}(\mathbf{c}./|\mathbf{c}^\star|)^H\mathbf{D}_2(\mathbf{f})/\tau\,\mathrm{diag}(\mathbf{c}./|\mathbf{c}^\star|)\} - \mathbf{I}\right\|_\infty + 2\left\|-\mathbb{R}\{\mathrm{diag}(\mathbf{c}./|\mathbf{c}^\star|)\}\mathbf{D}_1(\mathbf{f})/\sqrt{\tau}\right\|_\infty \\ &+ 2\left\|\mathrm{diag}(1./|\mathbf{c}^\star|)[\mathbf{D}_1(\mathbf{f}^\lambda,\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f})\mathbf{c}]/\sqrt{\tau}\right\|_\infty + \left\|\mathrm{diag}(\mathbf{c}./|\mathbf{c}^\star|^2)[\mathbf{D}_2(\mathbf{f}^\lambda,\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_2(\mathbf{f})\mathbf{c}]/\tau\right\|_\infty \\ \overset{①}{\leq}& (0.05610 + 2.12X^\star\gamma) + 2(1+X^\star\gamma)(0.01236) + 2(0.75038B^\star X^\star\gamma) + 1.14168B^\star X^\star\gamma \\ \leq& 0.08561, \end{aligned} \tag{A.39}$$

where ① follows from Eq. (A.11), (A.19), (A.20) and the following bound

$$\begin{aligned} \left\|-\mathbb{R}\{\mathrm{diag}(\mathbf{c}./|\mathbf{c}^\star|)^H\mathbf{D}_2(\mathbf{f})/\tau\,\mathrm{diag}(\mathbf{c}./|\mathbf{c}^\star|)\} - \mathbf{I}\right\|_\infty \leq& \max_i \left|\frac{|c_i|^2}{|c_i^\star|^2} - 1\right| + 0.05610\max_{i,j}\frac{|c_i||c_j|}{|c_i^\star||c_j^\star|} \\ \leq& X^\star\gamma(2+X^\star\gamma) + 0.05610(1+X^\star\gamma)^2 \\ \leq& 1.05610(X^\star\gamma)^2 + 2.113X^\star\gamma + 0.05610 \\ \leq& 0.05610 + 2.12X^\star\gamma. \end{aligned} \tag{A.40}$$

Bounding Π_2^λ and Π_3^λ.

Note Π_2^λ and Π_3^λ are of the same form. Thus we can bound them together:

$$\begin{aligned} \max\{\Pi_2^\lambda, \Pi_3^\lambda\} \leq& \|\mathbf{D}_1(\mathbf{f})\mathbb{R}\{\mathrm{diag}(\mathbf{c}./|\mathbf{c}^\star|)\}\|_\infty/\sqrt{\tau} + \|\mathrm{diag}(1./|\mathbf{c}^\star|)[\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f})\mathbf{c}]\|_\infty/\sqrt{\tau} \\ &+ \|\mathbf{D}_0(\mathbf{f}) - \mathbf{I}\|_\infty + 2\|\lambda\,\mathrm{diag}(\mathbf{u}\odot\mathbf{v}./\left|\mathbf{c}\right|^3)\|_\infty \\ \overset{①}{\leq}& (1+X^\star\gamma)(0.01236) + (0.75038B^\star X^\star\gamma) + (0.00755) + 2(0.646X^\star\gamma) \\ <& \Pi_1^\lambda \text{ (since } B^\star X^\star\gamma \leq 10^{-3}), \end{aligned}$$

where ① follows from Eq. (A.11), (A.19)- (A.20) and $\lambda \leq 0.646 X^\star \gamma_0$. Therefore,

$$\underset{\boldsymbol{\theta}\in\mathcal{N}^\star}{\text{maximize}} \left\| \Upsilon(\nabla^2 G^\lambda(\boldsymbol{\theta})) - \mathbf{I} \right\|_{\infty,\infty} \leq 0.08561 < 1, \tag{A.41}$$

implying the contraction property of $\Theta^\lambda(\boldsymbol{\theta})$.

A.4.2 Showing the Non-escaping Property

By the definition of the neighborhood $\mathcal{N}^\star$, it suffices to bound the distance between $\Theta^\lambda(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^\star$:

$$\begin{aligned}
\|\Theta^\lambda(\boldsymbol{\theta}) - \boldsymbol{\theta}^\star\|_\infty &\overset{①}{\leq} \|\Theta^\lambda(\boldsymbol{\theta}) - \Theta^\lambda(\boldsymbol{\theta}^\star)\|_\infty + \|\Theta^\lambda(\boldsymbol{\theta}^\star) - \boldsymbol{\theta}^\star\|_\infty \\
&\overset{②}{=} \| \int_0^1 [\nabla_{\boldsymbol{\theta}} \Theta^\lambda((1-t)\boldsymbol{\theta}^\star + t\boldsymbol{\theta})](\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\, \mathrm{d}\, t\|_\infty + \|\Theta^\lambda(\boldsymbol{\theta}^\star) - \boldsymbol{\theta}^\star\|_\infty \\
&\overset{③}{\leq} \underset{\mathbf{z}\in\mathcal{N}^\star}{\text{maximize}} \|\nabla_{\boldsymbol{\theta}} \Theta^\lambda(\mathbf{z})\|_{\infty,\infty} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_\infty + \|\mathbf{W}^\star \nabla G^\lambda(\boldsymbol{\theta}^\star)\|_\infty \\
&\overset{④}{\leq} (0.08561)(X^\star \gamma_0/\sqrt{2}) + \lambda \overset{⑤}{\leq} X^\star \gamma_0/\sqrt{2},
\end{aligned}$$

where ① follows from the triangle inequality, ② follows from the integral form of the mean value theorem for vector-valued functions (see [148, Eq. (A.57)]), ③ follows from sub-multiplicative property of $\|\cdot\|_{\infty,\infty}$ and the fact that $(1-t)\boldsymbol{\theta}^\star + t\boldsymbol{\theta}) \in \mathcal{N}^\star$ for $t \in [0,1]$, ④ follows from

$$\|\mathbf{W}^\star \nabla G^\lambda(\boldsymbol{\theta}^\star)\|_\infty = \left\| \begin{bmatrix} 0 \\ \mathbb{R}\{\lambda \mathbf{c}^\star ./ |\mathbf{c}^\star|\} \\ \mathbb{I}\{\lambda \mathbf{c}^\star ./ |\mathbf{c}^\star|\} \end{bmatrix} \right\|_\infty \leq \lambda,$$

and ⑤ holds for $\lambda \leq 0.646 X^\star \gamma_0$ since $(0.08561)(X^\star \gamma_0/\sqrt{2}) + 0.646 X^\star \gamma_0 \leq 0.9992 X^\star \gamma_0/\sqrt{2}$.

In sum, Θ^λ satisfies both the contraction and the non-escaping properties in $\mathcal{N}^\star$. Therefore, by the contraction mapping theorem, the map Θ^λ has a unique fixed point $\boldsymbol{\theta}^\lambda \in \mathcal{N}^\star$ satisfying $\Theta^\lambda(\boldsymbol{\theta}^\lambda) = \boldsymbol{\theta}^\lambda$.

We continue to show that $\boldsymbol{\theta}^\lambda$ is a differentiable function of λ. Define a function $F : \mathbb{R}^{3k} \times \mathbb{R} \mapsto \mathbb{R}^{3k}$ as $F(\boldsymbol{\theta}, \lambda) = \nabla G^\lambda(\boldsymbol{\theta})$ and recognize $F(\boldsymbol{\theta}, \lambda)$ is continuously differentiable since it has a continuous Jacobian given by

$$\partial F(\boldsymbol{\theta}, \lambda) = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\theta}} F(\boldsymbol{\theta}, \lambda) & \frac{\partial}{\partial \lambda} F(\boldsymbol{\theta}, \lambda) \end{bmatrix} = \begin{bmatrix} \nabla^2 G^\lambda(\boldsymbol{\theta}) & \begin{bmatrix} \mathbf{0} \\ \mathbb{R}\{\mathbf{c}./|\mathbf{c}|\} \\ \mathbb{I}\{\mathbf{c}./|\mathbf{c}|\} \end{bmatrix} \end{bmatrix},$$

with $\frac{\partial}{\partial \boldsymbol{\theta}} F(\boldsymbol{\theta}, \lambda)$ nonsingular in $\mathcal{N}^\star$ by (A.41). Then according to the implicit function theorem (see [269, Proposition A.25]), there is a continuously differentiable function $\mathbf{g}(\cdot)$ such that $F(\mathbf{g}(\lambda), \lambda) = \nabla G^\lambda(\mathbf{g}(\lambda)) = \mathbf{0}$ and

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \mathbf{g}(\lambda) = -(\frac{\partial}{\partial \boldsymbol{\theta}} F(\mathbf{g}(\lambda), \lambda))^{-1} \frac{\partial}{\partial \lambda} F(\mathbf{g}(\lambda), \lambda) = -(\nabla^2 G^\lambda(\mathbf{g}(\lambda)))^{-1} \frac{\partial}{\partial \lambda} \nabla G^\lambda(\mathbf{g}(\lambda)). \tag{A.42}$$

Since $\nabla G^\lambda(\mathbf{g}(\lambda)) = \mathbf{0}$ is equivalent to $\Theta^\lambda(\mathbf{g}(\lambda)) = \mathbf{g}(\lambda)$, we conclude that $\boldsymbol{\theta}^\lambda = \mathbf{g}(\lambda)$ due to the uniqueness of the fixed point of Θ^λ. Therefore, $\boldsymbol{\theta}^\lambda$ is a differentiable function of λ and

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\boldsymbol{\theta}^\lambda = -(\nabla^2 G^\lambda(\boldsymbol{\theta}^\lambda))^{-1}\frac{\partial}{\partial\lambda}\nabla G^\lambda(\boldsymbol{\theta}^\lambda). \tag{A.43}$$

Finally, let $\lim_{\lambda\to 0}\boldsymbol{\theta}^\lambda = \boldsymbol{\theta}^0$. Taking limit as λ goes to 0 in the equation $\nabla G^\lambda(\boldsymbol{\theta}^\lambda) = \mathbf{0}$ yields $\nabla G^0(\boldsymbol{\theta}^0) = \mathbf{0}$ due to the continuity of $\nabla G^\lambda(\boldsymbol{\theta})$ in λ and $\boldsymbol{\theta}$ and the continuity of $\boldsymbol{\theta}^\lambda$. Since $\nabla G^0(\boldsymbol{\theta}^\star) = \mathbf{0}$ by direct computation and the solution is unique in $\mathcal{N}^\star$, we conclude that $\lim_{\lambda\to 0}\boldsymbol{\theta}^\lambda = \boldsymbol{\theta}^0 = \boldsymbol{\theta}^\star$. □

A.5 Proof of Lemma 2.4.2

Lemma A.5.1 (Lemma 2.4.2)**.** *Let the second fixed point map be the weighted gradient map of the nonconvex program* (2.15)*:*

$$\Theta(\boldsymbol{\theta}) = \boldsymbol{\theta} - \mathbf{W}^\star\nabla\left(\frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{y}\|_{\mathbf{Z}}^2 + \lambda\|\mathbf{c}\|_1\right) \tag{2.21}$$

and the region $\mathcal{N}^\lambda := \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\lambda\|_\infty \le 35.2\gamma_0/\sqrt{2}\right\}$*. Set the regularization parameter* λ *as* $0.646X^\star\gamma_0$ *in* (2.21)*. Suppose that the separation condition* (2.9) *and the SNR condition* (2.10) *hold. Then with probability at least* $1 - \frac{1}{n^2}$*,* $\Theta(\boldsymbol{\theta})$ *has a unique fixed point* $\hat{\boldsymbol{\theta}}$ *living in* $\mathcal{N}^\lambda$*.*

Proof. The main idea is again to apply the contraction mapping theorem 2.4.1 to the fixed point map:

$$\Theta(\boldsymbol{\theta}) = \boldsymbol{\theta} - \mathbf{W}^\star\nabla G(\boldsymbol{\theta}),$$

where G is the objective function of (2.15):

$$G(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{A}(\mathbf{f})\mathbf{c} - \mathbf{y}\|_{\mathbf{Z}}^2 + \lambda\|\mathbf{c}\|_1$$

with $\lambda = 0.646X^\star\gamma_0$. By Theorem 2.4.1, showing the existence of a unique point $\hat{\boldsymbol{\theta}} \in \mathcal{N}^\lambda$ such that $\Theta(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}$ can be reduced to showing that Θ satisfies both the non-escaping property and the contraction properties:

(i) $\Theta(\mathcal{N}^\lambda) \subset \mathcal{N}^\lambda$;

(ii) There exists $\rho \in (0,1)$ such that $\|\Theta(\mathbf{v}) - \Theta(\mathbf{w})\|_\infty \le \rho\|\mathbf{v} - \mathbf{w}\|_\infty$ for any $\mathbf{v}, \mathbf{w} \in \mathcal{N}^\lambda$.

A.5.1 Showing the Contraction Property

Recall that $\mathcal{N}^\star$ is a neighborhood centered at $\boldsymbol{\theta}^\star$ and $\mathcal{N}^\lambda$ is a neighborhood centered at $\boldsymbol{\theta}^\lambda$ defined respectively via

$$\mathcal{N}^\star = \left\{ \boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_\infty \le \frac{X^\star}{\sqrt{2}}\gamma_0 \right\} \quad \text{and} \quad \mathcal{N}^\lambda = \left\{ \boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\lambda\|_\infty \le \frac{35.2}{\sqrt{2}}\gamma_0 \right\}.$$

Keep in mind that $\boldsymbol{\theta}^\lambda$ is the unique point in $\mathcal{N}^\star$ that satisfies $\nabla G^\lambda(\boldsymbol{\theta}^\lambda) = \mathbf{0}$. To show the contraction of Θ in $\mathcal{N}^\lambda$, our strategy is to show Θ is contractive in a larger set $\hat{\mathcal{N}}$ that contains $\mathcal{N}^\lambda$:

$$\hat{\mathcal{N}} = \left\{ \boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_\infty \le \frac{X^\star + 35.2}{\sqrt{2}}\gamma_0 := \frac{\hat{X}}{\sqrt{2}}\gamma_0 \right\}.$$

Recognize that $\hat{\mathcal{N}}$ is a neighborhood centered at $\boldsymbol{\theta}^\star$ but with a radius $35.2\gamma_0/\sqrt{2}$ larger than that of $\mathcal{N}^\star$. Such a choice is made for the purpose of showing the closeness between the final fixed point solution $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\star$. We remark that the quantity $35.2\gamma_0/\sqrt{2}$ corresponds to the dual atomic norm of the weighted Gaussian noise. Adding such a noise norm term to the radius of the original neighborhood $\mathcal{N}^\star$ ensures that the region $\hat{\mathcal{N}}$ is large enough for $\Theta(\boldsymbol{\theta})$ to be non-escaping. This is reasonable because the second fixed point map (2.21) involves an additive Gaussian noise and we have shown that the first fixed point map (2.19) (the one constructed in the noise-free setting) satisfies the non-escaping property in $\mathcal{N}^\star$.

Next, we apply arguments similar to those of showing the contraction of Θ^λ in $\mathcal{N}^\star$. In particular, we first compute the expression of $\Upsilon(\nabla^2 G(\boldsymbol{\theta})) - \mathbf{I}$:

$$\Upsilon(\nabla^2 G(\boldsymbol{\theta})) - \mathbf{I} = \begin{bmatrix} \frac{-1}{\tau}\mathbb{R}\{\boldsymbol{\Phi}^H \mathbf{D}_2(\mathbf{f})\boldsymbol{\Phi}\} - \mathbf{I} & \frac{-1}{\sqrt{\tau}}\mathbb{R}\{\boldsymbol{\Phi}\}\mathbf{D}_1(\mathbf{f}) & \frac{-1}{\sqrt{\tau}}\mathbb{I}\{\boldsymbol{\Phi}\}\mathbf{D}_1(\mathbf{f}) \\ \frac{1}{\sqrt{\tau}}\mathbf{D}_1(\mathbf{f})\mathbb{R}\{\boldsymbol{\Phi}\} & \mathbf{D}_0(\mathbf{f}) - \mathbf{I} & \mathbf{0} \\ \frac{1}{\sqrt{\tau}}\mathbf{D}_1(\mathbf{f})\mathbb{I}\{\boldsymbol{\Phi}\} & \mathbf{0} & \mathbf{D}_0(\mathbf{f}) - \mathbf{I} \end{bmatrix} + \begin{bmatrix} \operatorname{diag}(\hat{\mathbf{d}}_{\mathbf{ff}}) & \operatorname{diag}(\hat{\mathbf{d}}_{\mathbf{fu}}) & \operatorname{diag}(\hat{\mathbf{d}}_{\mathbf{fv}}) \\ \operatorname{diag}(\hat{\mathbf{d}}_{\mathbf{fu}}) & \operatorname{diag}(\hat{\mathbf{d}}_{\mathbf{uu}}) & \operatorname{diag}(\hat{\mathbf{d}}_{\mathbf{uv}}) \\ \operatorname{diag}(\hat{\mathbf{d}}_{\mathbf{fv}}) & \operatorname{diag}(\hat{\mathbf{d}}_{\mathbf{uv}}) & \operatorname{diag}(\hat{\mathbf{d}}_{\mathbf{vv}}) \end{bmatrix}$$

with $\boldsymbol{\Phi} = \operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|)$ and

$$\begin{aligned}
\hat{\mathbf{d}}_{\mathbf{ff}} &= -\mathbb{R}\{\operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|^2)^H[\mathbf{A}''(\mathbf{f})^H\mathbf{Z}\mathbf{w} + \mathbf{D}_2(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_2(\mathbf{f})\mathbf{c}]/\tau\}; \quad & \hat{\mathbf{d}}_{\mathbf{uu}} &= \lambda \operatorname{diag}(\mathbf{u}\odot\mathbf{u}./\left|\mathbf{c}\right|^3); \\
\hat{\mathbf{d}}_{\mathbf{fu}} &= \mathbb{R}\{\operatorname{diag}(1./|\mathbf{c}^\star|)[-\mathbf{A}'(\mathbf{f})^H\mathbf{Z}\mathbf{w} + \mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f})\mathbf{c}]/\sqrt{\tau}\}; \quad & \hat{\mathbf{d}}_{\mathbf{uv}} &= \lambda \operatorname{diag}(\mathbf{u}\odot\mathbf{v}./\left|\mathbf{c}\right|^3); \\
\hat{\mathbf{d}}_{\mathbf{fv}} &= \mathbb{I}\{\operatorname{diag}(1./|\mathbf{c}^\star|)[-\mathbf{A}'(\mathbf{f})^H\mathbf{Z}\mathbf{w} + \mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f})\mathbf{c}]/\sqrt{\tau}\}; \quad & \hat{\mathbf{d}}_{\mathbf{vv}} &= \lambda \operatorname{diag}(\mathbf{v}\odot\mathbf{v}./\left|\mathbf{c}\right|^3).
\end{aligned}$$

Comparing the expressions for $[\Upsilon(\nabla^2 G^\lambda(\boldsymbol{\theta})) - \mathbf{I}]$ and $[\Upsilon(\nabla^2 G(\boldsymbol{\theta})) - \mathbf{I}]$ shows that the latter differs in have additional noise terms in the first row and the first column blocks. We have shown that the first absolute row sum Π_1^λ of $[\Upsilon(\nabla^2 G^\lambda(\boldsymbol{\theta})) - \mathbf{I}]$ dominates the other row sums. Having additional noise terms will only increase the final bounds due to the application of the triangle inequality. Therefore, the first absolute row sum (denoted by $\hat{\Pi}_1$) of $[\Upsilon(\nabla^2 G(\boldsymbol{\theta})) - \mathbf{I}]$ also dominates and hence achieves the $\ell_{\infty,\infty}$ norm. Direct computation gives

$$\begin{aligned}
\hat{\Pi}_1 &\le \Pi_1^\lambda + 2\|\operatorname{diag}(1./|\mathbf{c}^\star|)\mathbf{A}'(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty/\sqrt{\tau} + \|\operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|^2)\mathbf{A}''(\mathbf{f})^H\mathbf{Z}\mathbf{w}\|_\infty/\tau \\
&\overset{①}{\le} 0.08561 + 2(22.64\gamma) + 78.43(1 + \hat{X}\gamma)\gamma \\
&\overset{②}{\le} 0.08563,
\end{aligned}$$

where ① follows from $\Pi_1^\lambda \leq 0.08561$ and Eq. (A.32)- (A.34), ② follows from $\hat{X} = X^\star + 35.2$ and the SNR condition (2.10) that $X^\star B^\star \gamma \leq 10^{-3}$ and $B^\star/X^\star \leq 10^{-4}$ hence $2(22.64\gamma) + 78.43(1+\hat{X}\gamma)\gamma \leq 0.00002$. Hence,

$$\underset{\boldsymbol{\theta}\in\hat{\mathcal{N}}}{\text{maximize}} \|\nabla\Theta(\boldsymbol{\theta})\|_{\hat{\infty},\hat{\infty}} \leq 0.08563 < 1. \tag{A.44}$$

This implies the contraction of Θ in $\mathcal{N}^\lambda$, since

$$\underset{\boldsymbol{\theta}\in\mathcal{N}^\lambda}{\text{maximize}} \|\nabla\Theta(\boldsymbol{\theta})\|_{\hat{\infty},\hat{\infty}} \leq \underset{\boldsymbol{\theta}\in\hat{\mathcal{N}}}{\text{maximize}} \|\nabla\Theta(\boldsymbol{\theta})\|_{\hat{\infty},\hat{\infty}}.$$

A.5.2 Showing the Non-escaping Property

$$\begin{aligned}
\|\Theta(\boldsymbol{\theta}) - \boldsymbol{\theta}^\lambda\|_{\hat{\infty}} =& \|(\Theta(\boldsymbol{\theta}) - \Theta(\boldsymbol{\theta}^\lambda)) + (\Theta(\boldsymbol{\theta}^\lambda) - \boldsymbol{\theta}^\lambda)\|_{\hat{\infty}} \\
\overset{①}{\leq}& \|\nabla\Theta(\tilde{\boldsymbol{\theta}})^T(\boldsymbol{\theta} - \boldsymbol{\theta}^\lambda)\|_{\hat{\infty}} + \|\mathbf{W}^\star\nabla G(\boldsymbol{\theta}^\lambda)\|_{\hat{\infty}} \\
\leq& \max_{\tilde{\boldsymbol{\theta}}\in\hat{\mathcal{N}}} \|\nabla\Theta(\tilde{\boldsymbol{\theta}})\|_{\hat{\infty},\hat{\infty}} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\lambda\|_{\hat{\infty}} + \|\mathbf{W}^\star\nabla G(\boldsymbol{\theta}^\lambda)\|_{\hat{\infty}} \\
\overset{②}{\leq}& (0.08563)\left(35.2\gamma_0/\sqrt{2}\right) + 22.7\gamma_0 \\
\leq& 35.117\gamma_0/\sqrt{2} < 35.2\gamma_0/\sqrt{2},
\end{aligned}$$

where ① follows from the mean value theorem for some $\tilde{\boldsymbol{\theta}}$ on the line segment joining $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\lambda$ and ② follows from (A.44) and (A.45). Eq. (A.45) is given as follows

$$\begin{aligned}
\left\|\mathbf{W}^\star\nabla G(\boldsymbol{\theta}^\lambda)\right\|_{\hat{\infty}} =& \left\| \mathbf{W}^\star \begin{bmatrix} \mathbb{R}\{-\operatorname{diag}(\mathbf{c}^\lambda)^H(\mathbf{A}'(\mathbf{f}^\lambda)^H\mathbf{Z}\mathbf{w} + \mathbf{D}_1(\mathbf{f}^\lambda,\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f}^\lambda)\mathbf{c}^\lambda)\} \\ \mathbb{R}\{-\mathbf{A}(\mathbf{f}^\lambda)^H\mathbf{Z}\mathbf{w} - \mathbf{D}_0(\mathbf{f}^\lambda,\mathbf{f}^\star)\mathbf{c}^\star + \mathbf{D}_0(\mathbf{f}^\lambda)\mathbf{c}^\lambda + \lambda\mathbf{c}^\lambda./|\mathbf{c}^\lambda|\} \\ \mathbb{I}\{-\mathbf{A}(\mathbf{f}^\lambda)^H\mathbf{Z}(^\lambda\mathbf{w} - \mathbf{D}_0(\mathbf{f}^\lambda,\mathbf{f}^\star)\mathbf{c}^\star + \mathbf{D}_0(\mathbf{f}^\lambda)\mathbf{c}^\lambda + \lambda\mathbf{c}^\lambda./|\mathbf{c}^\lambda|\} \end{bmatrix} \right\|_{\hat{\infty}} \\
\overset{①}{=}& \left\| \begin{bmatrix} \mathbb{R}\{-\operatorname{diag}(\mathbf{c}^\lambda./|\mathbf{c}^\star|)^H\mathbf{A}'(\mathbf{f}^\lambda)^H\mathbf{Z}\mathbf{w}\}/\sqrt{\tau} \\ \mathbb{R}\{-\mathbf{A}(\mathbf{f}^\lambda)^H\mathbf{Z}\mathbf{w}\} \\ \mathbb{I}\{-\mathbf{A}(\mathbf{f}^\lambda)^H\mathbf{Z}\mathbf{w}\} \end{bmatrix} \right\|_\infty \\
\overset{②}{\leq}& \left\| \begin{bmatrix} 41.052n/\sqrt{\tau}(1+X^\star\gamma)\gamma_0 \\ 6.534\gamma_0 \\ 6.534\gamma_0 \end{bmatrix} \right\|_\infty \\
\leq& 22.7\gamma_0,
\end{aligned} \tag{A.45}$$

where ① holds since $\nabla G^\lambda(\boldsymbol{\theta})$ vanishes at $\boldsymbol{\theta}^\lambda$:

$$\nabla G^\lambda(\boldsymbol{\theta}^\lambda) = \begin{bmatrix} \mathbb{R}\{-\operatorname{diag}(\mathbf{c}^\lambda)^H(\mathbf{D}_1(\mathbf{f}^\lambda,\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f}^\lambda)\mathbf{c}^\lambda)\} \\ \mathbb{R}\{-\mathbf{D}_0(\mathbf{f}^\lambda,\mathbf{f}^\star)\mathbf{c}^\star + \mathbf{D}_0(\mathbf{f}^\lambda)\mathbf{c}^\lambda + \lambda\mathbf{c}^\lambda./|\mathbf{c}^\lambda|\} \\ \mathbb{I}\{-\mathbf{D}_0(\mathbf{f}^\lambda,\mathbf{f}^\star)\mathbf{c}^\star + \mathbf{D}_0(\mathbf{f}^\lambda)\mathbf{c}^\lambda + \lambda\mathbf{c}^\lambda./|\mathbf{c}^\lambda|\} \end{bmatrix} = \mathbf{0}.$$

② holds with probability at least $1 - \frac{1}{n^2}$ by (A.32)- (A.34).

Hence both the contraction and the non-escaping properties are satisfied by Θ in $\mathcal{N}^\lambda$. Then by the contraction mapping theorem, we conclude the proof of Lemma 2.4.2. □

A.6 Proof of Lemma 2.4.3

Lemma A.6.1 (Lemma 2.4.3)**.** *The dual polynomial $Q^\star(f)$ satisfies both the Interpolation and Boundedness properties with respect to the coefficients $\{c_\ell^\star\}$ and the frequencies $\{f_\ell^\star\}$. In addition, $Q^\star(f)$ satisfies first*

$$\begin{array}{ll} Q_R^\star(f) \geq 0.887594, & Q_R^\star{}''(f) \leq -2.24483n^2, \\ |Q_I^\star(f)| \leq 0.0183836, & |Q^\star{}_I''(f)| \leq 0.113197n^2, \\ |Q^{\star\prime}(f)| \leq 0.821039n, & |Q^{\star\prime\prime}(f)| \leq 3.40320n^2, \end{array}$$

and

$$Q_R^\star(f)Q_R^\star(f)'' + |Q^\star(f)'|^2 + |Q_I^\star(f)||Q_I^\star(f)''| \leq -1.316313n^2 < 0$$

for $f \in \mathcal{N}$, implying $|Q^\star(f)|'' < 0$ in $\mathcal{N}$, and second,

$$\begin{array}{l} |Q^\star(f)| \leq 0.927615,\ f \in \mathcal{M}, \\ |Q^\star(f)| \leq 0.734123,\ f \in \mathcal{F}. \end{array}$$

Here the subscripts R and I denote respectively the real and imaginary parts of $Q^\star(f)$. Thus $\mathbf{q}^\star$ is a valid dual certificate to certify the atomic decomposition $\mathbf{x}^\star = \sum_{\ell=1}^k c_\ell^\star \mathbf{a}(f_\ell^\star)$ such that $\|\mathbf{x}^\star\|_{\mathcal{A}} = \sum_{\ell=1}^k |c_\ell^\star|$.

Proof. To show that $\mathbf{q}^\star$ is a valid dual certificate, it is instructive to first relate $\mathbf{q}^\star$ to the derivative of $\mathbf{x}^\lambda$ with respect to λ (where we treat $\mathbf{x}^\lambda$ as a function of λ):

$$\mathbf{q}^\star = \lim_{\lambda\to 0} \mathbf{q}^\lambda = \lim_{\lambda\to 0} \frac{\mathbf{x}^\star - \mathbf{x}^\lambda}{\lambda} = -\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathbf{x}^\lambda\big|_{\lambda=0}, \tag{A.46}$$

where we used the fact that $\lim_{\lambda\to 0}\mathbf{x}^\lambda = \lim_{\lambda\to 0}\mathbf{A}(\mathbf{f}^\lambda)\mathbf{c}^\lambda = \mathbf{A}(\mathbf{f}^\star)\mathbf{c}^\star = \mathbf{x}^\star$ by Lemma 2.4.1. Since $\mathbf{x}^\lambda = \mathbf{A}(\mathbf{f}^\lambda)\mathbf{c}^\lambda = \sum_\ell c_\ell^\lambda \mathbf{a}(f_\ell^\lambda)$, we compute the derivative $\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathbf{x}^\lambda$ using the chain rule as:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathbf{x}^\lambda = \sum_\ell \left(\frac{\mathrm{d}}{\mathrm{d}\lambda}u_\ell^\lambda + i\frac{\mathrm{d}}{\mathrm{d}\lambda}v_\ell^\lambda\right)\mathbf{a}(f_\ell^\lambda) + \sum_\ell c_\ell^\lambda\left(\frac{\mathrm{d}f_\ell^\lambda}{\mathrm{d}\lambda}\mathbf{a}'(f_\ell^\lambda)\right) = \left[\mathbf{A}'(\mathbf{f}^\lambda)\,\mathrm{diag}(\mathbf{c}^\lambda)\ \ \mathbf{A}(\mathbf{f}^\lambda)\ \ i\mathbf{A}(\mathbf{f}^\lambda)\right]\frac{\mathrm{d}}{\mathrm{d}\lambda}\boldsymbol{\theta}^\lambda, \tag{A.47}$$

where $\mathbf{A}'(\mathbf{f}) = \begin{bmatrix}\mathbf{a}'(f_1) & \cdots & \mathbf{a}'(f_k)\end{bmatrix}$. Therefore, using Eq. (A.46) and (A.47) we obtain:

$$\begin{aligned} \mathbf{q}^\star &= -\lim_{\lambda\to 0}\left[\mathbf{A}'(\mathbf{f}^\lambda)\,\mathrm{diag}(\mathbf{c}^\lambda)\ \ \mathbf{A}(\mathbf{f}^\lambda)\ \ i\mathbf{A}(\mathbf{f}^\lambda)\right]\frac{\mathrm{d}}{\mathrm{d}\lambda}\boldsymbol{\theta}^\lambda \\ &= -\left[\mathbf{A}'(\mathbf{f}^\star)\,\mathrm{diag}(\mathbf{c}^\star)\ \ \mathbf{A}(\mathbf{f}^\star)\ \ i\mathbf{A}(\mathbf{f}^\star)\right]\lim_{\lambda\to 0}\frac{\mathrm{d}}{\mathrm{d}\lambda}\boldsymbol{\theta}^\lambda \\ &= \left[\mathbf{A}'(\mathbf{f}^\star)\,\mathrm{diag}(\mathbf{c}^\star)\ \ \mathbf{A}(\mathbf{f}^\star)\ \ i\mathbf{A}(\mathbf{f}^\star)\right](\nabla^2 G^0(\boldsymbol{\theta}^\star))^{-1}\frac{\partial}{\partial\lambda}\nabla G^0(\boldsymbol{\theta}^\star), \end{aligned} \tag{A.48}$$

where in the second line we again used the fact that $\lim_{\lambda\to 0}\boldsymbol{\theta}^\lambda = \boldsymbol{\theta}^\star$ by Lemma 2.4.1, and in the last line we used the expression for $\mathrm{d}\boldsymbol{\theta}^\lambda/\mathrm{d}\lambda$ given in (2.20).

We next compute $\frac{\partial}{\partial\lambda}\nabla G^0(\boldsymbol{\theta}^\star)$ explicitly. Let $K^{(\ell)}(\cdot)$ denote the ℓ-order derivative of the Jackson kernel $K(\cdot)$ (see Appendix A.1 for more details). Recall that $\mathbf{D}_\ell(\mathbf{f}^1,\mathbf{f}^2) := [K^{(\ell)}(f_m^2 - f_n^1)]_{1\le n\le k, 1\le m\le k}$ and $\mathbf{D}_\ell(\mathbf{f}) := \mathbf{D}_\ell(\mathbf{f},\mathbf{f})$ are matrices formed by sampling the Jackson kernel and its derivatives. Then we have the following expression for $\nabla G^\lambda(\boldsymbol{\theta})$ (see Appendix A.3 for more details)

$$\nabla G^\lambda(\boldsymbol{\theta}) = \begin{bmatrix} \mathbb{R}\{\mathrm{diag}(\mathbf{c})(\mathbf{D}_1(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f})\mathbf{c})\} \\ \mathbb{R}\{-\mathbf{D}_0(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star + \mathbf{D}_0(\mathbf{f})\mathbf{c} + \lambda\mathbf{c}./|\mathbf{c}|\} \\ \mathbb{I}\{-\mathbf{D}_0(\mathbf{f},\mathbf{f}^\star)\mathbf{c}^\star + \mathbf{D}_0(\mathbf{f})\mathbf{c} + \lambda\mathbf{c}./|\mathbf{c}|\} \end{bmatrix}. \tag{A.49}$$

Therefore, the partial derivative of (A.49) with respect to λ is the expanded complex sign vector:

$$\frac{\partial}{\partial\lambda}\nabla G^\lambda(\boldsymbol{\theta}^\lambda) = \begin{bmatrix} \mathbf{0} \\ \mathbb{R}\{\mathrm{sign}(\mathbf{c}^\lambda)\} \\ \mathbb{I}\{\mathrm{sign}(\mathbf{c}^\lambda)\} \end{bmatrix} := \begin{bmatrix} \mathbf{0} \\ \mathbf{s}_R^\lambda \\ \mathbf{s}_I^\lambda \end{bmatrix} \quad\Longrightarrow\quad \frac{\partial}{\partial\lambda}\nabla G^0(\boldsymbol{\theta}^\star) = \begin{bmatrix} \mathbf{0} \\ \mathbb{R}\{\mathrm{sign}(\mathbf{c}^\star)\} \\ \mathbb{I}\{\mathrm{sign}(\mathbf{c}^\star)\} \end{bmatrix} := \begin{bmatrix} \mathbf{0} \\ \mathbf{s}_R^\star \\ \mathbf{s}_I^\star \end{bmatrix}. \tag{A.50}$$

Here $\mathbf{s}^\lambda = \mathbf{c}^\lambda./|\mathbf{c}^\lambda|, \mathbf{s}^\star = \mathbf{c}^\star./|\mathbf{c}^\star|$ and the subscript R and I indicate the real and imaginary parts respectively.

Combining Eq. (A.48) and (A.50), we get

$$\mathbf{q}^\star = [\mathbf{A}'(\mathbf{f}^\star)\ \mathbf{A}(\mathbf{f}^\star)\ i\mathbf{A}(\mathbf{f}^\star)] \underbrace{\begin{bmatrix} \mathrm{diag}(\mathbf{c}^\star) & & \\ & \mathbf{I} & \\ & & \mathbf{I} \end{bmatrix} (\nabla^2 G^0(\boldsymbol{\theta}^\star))^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{s}_R^\star \\ \mathbf{s}_I^\star \end{bmatrix}}_{:=[\boldsymbol{\beta}^T\ \boldsymbol{\alpha}_R^T\ \boldsymbol{\alpha}_I^T]^T}, \tag{A.51}$$

where we have defined the coefficient vectors $\boldsymbol{\alpha}_R, \boldsymbol{\alpha}_I$ and $\boldsymbol{\beta}$ in (A.51). These coefficient vectors satisfy

$$\nabla^2 G^0(\theta^\star) \begin{bmatrix} \mathrm{diag}(\mathbf{c}^\star)^{-1}\boldsymbol{\beta} \\ \boldsymbol{\alpha}_R \\ \boldsymbol{\alpha}_I \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{s}_R^\star \\ \mathbf{s}_I^\star \end{bmatrix}. \tag{A.52}$$

By denoting $\boldsymbol{\alpha} = \boldsymbol{\alpha}_R + i\boldsymbol{\alpha}_I$ and $\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_k]^T$, $\boldsymbol{\beta} = [\beta_1,\ldots,\beta_k]^T$, we obtain an explicit form for the dual polynomial $Q^\star(f)$:

$$Q^\star(f) = \mathbf{a}(f)^H\mathbf{Z}\mathbf{q}^\star = \sum_{\ell=1}^k \alpha_\ell K(f_\ell^\star - f) + \sum_{\ell=1}^k \beta_\ell K'(f_\ell^\star - f). \tag{A.53}$$

To show that $\mathbf{q}^\star$ certifies the atomic decomposition $\mathbf{x}^\star = \sum_{\ell=1}^k c_\ell^\star \mathbf{a}(f_\ell^\star)$, we need to establish that

1. $Q^\star(f)$ satisfies $Q^\star(f_\ell^\star) = \mathrm{sign}(c_\ell^\star), \ell = 1,\ldots,k$ (Interpolation);
2. $|Q^\star(f)| < 1, \forall f \notin T^\star$ (Boundedness).

A.6.1 Showing the Interpolation Property

The Interpolation property follows from the construction process and is also easy to verify directly by noting

$$\nabla^2 G^0(\boldsymbol{\theta}^\star) = \begin{bmatrix} -\mathbb{R}\{\operatorname{diag}(\mathbf{c}^\star)^H \mathbf{D}_2(\mathbf{f}^\star)\operatorname{diag}(\mathbf{c}^\star)\} & \mathbb{R}\{-\operatorname{diag}(\mathbf{c}^\star)^H \mathbf{D}_1(\mathbf{f}^\star)\} & \mathbb{I}\{\operatorname{diag}(\mathbf{c}^\star)^H \mathbf{D}_1(\mathbf{f}^\star)\} \\ -\mathbb{R}\{\mathbf{D}_1(\mathbf{f}^\star)^H \operatorname{diag}(\mathbf{c}^\star)\} & \mathbf{D}_0(\mathbf{f}^\star) & 0 \\ -\mathbb{I}\{\mathbf{D}_1(\mathbf{f}^\star)^H \operatorname{diag}(\mathbf{c}^\star)\} & 0 & \mathbf{D}_0(\mathbf{f}^\star) \end{bmatrix}.$$

Indeed, the Interpolation property is a result of (A.52): since $\mathbf{D}_1(\mathbf{f}^\star) \in \mathbb{R}^{k\times k}$ and $\mathbf{D}_1(\mathbf{f}^\star)^T = -\mathbf{D}_1(\mathbf{f}^\star)$ (see Appendix A.1), the last two row blocks in (A.52) read

$$\begin{aligned} &\begin{bmatrix} \mathbf{D}_1(\mathbf{f}^\star)\mathbb{R}\{\operatorname{diag}(\mathbf{c}^\star)\} & \mathbf{D}_0(\mathbf{f}^\star) & 0 \\ \mathbf{D}_1(\mathbf{f}^\star)\mathbb{I}\{\operatorname{diag}(\mathbf{c}^\star)\} & 0 & \mathbf{D}_0(\mathbf{f}^\star) \end{bmatrix} \begin{bmatrix} \operatorname{diag}(\mathbf{c}^\star)^{-1}\boldsymbol{\beta} \\ \boldsymbol{\alpha}_R \\ \boldsymbol{\alpha}_I \end{bmatrix} = \begin{bmatrix} \mathbf{s}_R^\star \\ \mathbf{s}_I^\star \end{bmatrix} \\ \iff\ & \mathbf{D}_1(\mathbf{f}^\star)(\mathbb{R}\{\operatorname{diag}(\mathbf{c}^\star)\} + i\mathbb{I}\{\operatorname{diag}(\mathbf{c}^\star)\})\operatorname{diag}(\mathbf{c}^\star)^{-1}\boldsymbol{\beta} + \mathbf{D}_0(\mathbf{f}^\star)(\boldsymbol{\alpha}_R + i\boldsymbol{\alpha}_I) = \mathbb{R}\{\operatorname{sign}(\mathbf{c}^\star)\} + i\mathbb{I}\{\operatorname{sign}(\mathbf{c}^\star)\} \\ \iff\ & \mathbf{D}_1(\mathbf{f}^\star)\boldsymbol{\beta} + \mathbf{D}_0(\mathbf{f}^\star)\boldsymbol{\alpha} = \operatorname{sign}(\mathbf{c}^\star) \\ \iff\ & Q^\star(f_\ell^\star) = \operatorname{sign}(c_\ell^\star), \ell = 1, \ldots, k. \end{aligned} \tag{A.54}$$

Furthermore, the first row block of (A.52) is equivalent to

$$\begin{aligned} &-\mathbb{R}\{\operatorname{diag}(\mathbf{c}^\star)^H \mathbf{D}_2(\mathbf{f}^\star)\operatorname{diag}(\mathbf{c}^\star)\}\operatorname{diag}(\mathbf{c}^\star)^{-1}\boldsymbol{\beta} + \mathbb{R}\{-\operatorname{diag}(\mathbf{c}^\star)^H \mathbf{D}_1(\mathbf{f}^\star)\}\boldsymbol{\alpha}_R + \mathbb{I}\{\operatorname{diag}(\mathbf{c}^\star)^H \mathbf{D}_1(\mathbf{f}^\star)\}\boldsymbol{\alpha}_I = \mathbf{0} \\ \iff\ & \mathbb{R}\{\operatorname{diag}(\mathbf{c}^\star)^H (\mathbf{D}_2(\mathbf{f}^\star)\boldsymbol{\beta} + \mathbf{D}_1(\mathbf{f}^\star)\boldsymbol{\alpha})\} = \mathbf{0} \\ \iff\ & \mathbb{R}\{c_\ell^{\star H} Q^\star(f_\ell)'\} = 0, \ell = 1, \ldots, k. \end{aligned} \tag{A.55}$$

A.6.2 Showing the Boundedness Property

It remains to show that $Q^\star(f)$ satisfies the Boundedness property, for which we follow the arguments of [13]. We start with estimating the coefficient vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by rewriting (A.51) as

$$\begin{bmatrix} \operatorname{diag}(\mathbf{c}^\star) & & \\ & \mathbf{I} & \\ & & \mathbf{I} \end{bmatrix} \Phi \left(\Phi \nabla^2 G^0(\boldsymbol{\theta}^\star)\Phi\right)^{-1} \Phi \begin{bmatrix} \mathbf{0} \\ \mathbf{s}_R^\star \\ \mathbf{s}_I^\star \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha}_R \\ \boldsymbol{\alpha}_I \end{bmatrix}, \tag{A.56}$$

where $\Phi = \operatorname{diag}\left(\left[\operatorname{diag}\left(\frac{1}{|\mathbf{c}^\star|}\right), \mathbf{I}, \mathbf{I}\right]\right)$. Denoting $\boldsymbol{\Phi} := \operatorname{diag}(\mathbf{s}^\star)$, we further simplify (A.56) as

$$\begin{bmatrix} -\mathbb{R}\{\boldsymbol{\Phi}^H \mathbf{D}_2(\mathbf{f}^\star)\boldsymbol{\Phi}\} & \mathbb{R}\{-\boldsymbol{\Phi}^H \mathbf{D}_1(\mathbf{f}^\star)\} & \mathbb{I}\{\boldsymbol{\Phi}^H \mathbf{D}_1(\mathbf{f}^\star)\} \\ -\mathbb{R}\{\mathbf{D}_1(\mathbf{f}^\star)^H \boldsymbol{\Phi}\} & \mathbf{D}_0(\mathbf{f}^\star) & 0 \\ -\mathbb{I}\{\mathbf{D}_1(\mathbf{f}^\star)^H \boldsymbol{\Phi}\} & 0 & \mathbf{D}_0(\mathbf{f}^\star) \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}^{-1}\boldsymbol{\beta} \\ \boldsymbol{\alpha}_R \\ \boldsymbol{\alpha}_I \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{s}_R^\star \\ \mathbf{s}_I^\star \end{bmatrix}. \tag{A.57}$$

Denote

$$\begin{aligned} \tilde{\mathbf{D}}_2 &= -\mathbb{R}\{\operatorname{diag}(\mathbf{s}^\star)^H \mathbf{D}_2(\mathbf{f}^\star)\operatorname{diag}(\mathbf{s}^\star)\}; \\ \tilde{\mathbf{D}}_1 &= \operatorname{diag}(\mathbf{s}^\star)^H \mathbf{D}_1(\mathbf{f}^\star); \\ \tilde{\boldsymbol{\beta}} &= \operatorname{diag}(\mathbf{s}^\star)^{-1}\boldsymbol{\beta}. \end{aligned}$$

The last two row blocks of (A.57) give

$$\begin{aligned}\boldsymbol{\alpha}_R &= \mathbf{D}_0(\mathbf{f}^\star)^{-1}[\mathbf{s}_R^\star + \mathbb{R}\{\mathbf{D}_1(\mathbf{f}^\star)^H \operatorname{diag}(\mathbf{s}^\star)\}\tilde{\boldsymbol{\beta}}];\\ \boldsymbol{\alpha}_I &= \mathbf{D}_0(\mathbf{f}^\star)^{-1}[\mathbf{s}_I^\star + \mathbb{I}\{\mathbf{D}_1(\mathbf{f}^\star)^H \operatorname{diag}(\mathbf{s}^\star)\}\tilde{\boldsymbol{\beta}}]\end{aligned}$$

implying

$$\begin{aligned}\boldsymbol{\alpha} &= \mathbf{D}_0(\mathbf{f}^\star)^{-1}[\mathbf{s}^\star + \mathbf{D}_1(\mathbf{f}^\star)^H \operatorname{diag}(\mathbf{s}^\star)\tilde{\boldsymbol{\beta}}] \\ &= \mathbf{D}_0(\mathbf{f}^\star)^{-1}[\mathbf{s}^\star + \mathbf{D}_1(\mathbf{f}^\star)^H\boldsymbol{\beta}] \\ &= \mathbf{s}^\star - (\mathbf{I} - \mathbf{D}_0(\mathbf{f}^\star)^{-1})\mathbf{s}^\star + \mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{D}_1(\mathbf{f}^\star)^H\boldsymbol{\beta}.\end{aligned} \tag{A.58}$$

Without loss of generality, we assume $\mathbf{e}_1^T\mathbf{s}^\star = 1$. Then

$$\alpha_1 = 1 - \left[(\mathbf{I} - \mathbf{D}_0(\mathbf{f}^\star)^{-1})\mathbf{s}^\star - \mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{D}_1(\mathbf{f}^\star)^H\boldsymbol{\beta}\right]_1, \tag{A.59}$$

where $[\cdot]_1$ stands for the first entry of a vector. The first row block of (A.57) leads to

$$\tilde{\mathbf{D}}_2\tilde{\boldsymbol{\beta}} = \mathbb{R}\{\tilde{\mathbf{D}}_1\boldsymbol{\alpha}_R\} - \mathbb{I}\{\tilde{\mathbf{D}}_1\boldsymbol{\alpha}_I\} = \mathbb{R}\{\tilde{\mathbf{D}}_1(\boldsymbol{\alpha}_R + i\boldsymbol{\alpha}_I)\} = \mathbb{R}\{\tilde{\mathbf{D}}_1\boldsymbol{\alpha}\}.$$

Combining this with (A.58), we get

$$\begin{aligned}\tilde{\mathbf{D}}_2\tilde{\boldsymbol{\beta}} &= \mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}[\mathbf{s}^\star + \mathbf{D}^{\star H}_1\boldsymbol{\beta}]\} \\ &= \mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{s}^\star\} + \mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\}\mathbf{D}^{\star H}_1\boldsymbol{\beta} \\ &= \mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{s}^\star\} + \mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\}\mathbf{D}^{\star H}_1 \operatorname{diag}(\mathbf{s}^\star)\tilde{\boldsymbol{\beta}} \\ &= \mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{s}^\star\} + \mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\}\tilde{\mathbf{D}}^H_1\tilde{\boldsymbol{\beta}}.\end{aligned}$$

This implies

$$(\tilde{\mathbf{D}}_2 - \mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\}\tilde{\mathbf{D}}^H_1)\tilde{\boldsymbol{\beta}} = \mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{s}^\star\}. \tag{A.60}$$

A.6.2.1 Bounding $\|\tilde{\boldsymbol{\beta}}\|_\infty$

First invoke (A.11) to get

$$\begin{aligned}\|\mathbf{D}_0(\mathbf{f}^\star)^{-1}\|_{\infty,\infty} &\le \frac{1}{1-0.00755}, \\ \{\|\mathbf{D}_1(\mathbf{f}^\star)\|_{\infty,\infty}, \|\tilde{\mathbf{D}}_1\|_{\infty,\infty}\}/\sqrt{\tau} &\le 0.01236n/\sqrt{\tau} \le 0.00682, \\ \|\mathbf{I} - \tilde{\mathbf{D}}_2/\tau\|_{\infty,\infty} &\le 0.0171.\end{aligned} \tag{A.61}$$

These inequalities (A.61) immediately lead to

$$\begin{aligned}\|\tau\mathbf{I}-\tilde{\mathbf{D}}_2+\mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\}\tilde{\mathbf{D}}_1^H\|_{\infty,\infty}\overset{①}{\leq}&\tau\left(\|\mathbf{I}-\tilde{\mathbf{D}}_2/\tau\|_{\infty,\infty}+\|\tilde{\mathbf{D}}_1/\sqrt{\tau}\|_{\infty,\infty}^2\|\mathbf{D}_0(\mathbf{f}^\star)^{-1}\|_{\infty,\infty}\right)\\ \overset{②}{\leq}&\tau\left(0.0171+0.00682^2/(1-0.00755)\right)\leq 0.01715\tau<\tau,\end{aligned}\tag{A.62}$$

where ① follows from the triangle inequality and the sub-multiplicative property of $\ell_{\infty,\infty}$ norm and ② follows from (A.61). This implies that $\tilde{\mathbf{D}}_2-\mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\}\tilde{\mathbf{D}}_1^H$ is nonsingular and well-conditioned. In particular,

$$\|(\tilde{\mathbf{D}}_2-\mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\}\tilde{\mathbf{D}}_1^H)^{-1}\|_{\infty,\infty}\overset{①}{\leq}\frac{1}{\tau(1-0.01715)}\leq\frac{1.0175}{\tau},$$

where ① follows from (A.62). Then from (A.60), we have

$$\begin{aligned}\left\|\tilde{\boldsymbol{\beta}}\right\|_\infty\leq&\|(\tilde{\mathbf{D}}_2-\mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\}\tilde{\mathbf{D}}_1^H)^{-1}\|_{\infty,\infty}\|\mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{s}^\star\}\|_\infty\\ \overset{①}{\leq}&\|(\tilde{\mathbf{D}}_2-\mathbb{R}\{\tilde{\mathbf{D}}_1\mathbf{D}_0(\mathbf{f}^\star)^{-1}\}\tilde{\mathbf{D}}_1^H)^{-1}\|_{\infty,\infty}\|\tilde{\mathbf{D}}_1\|_{\infty,\infty}\|\mathbf{D}_0(\mathbf{f}^\star)^{-1}\|_{\infty,\infty}\|\mathbf{s}^\star\|_\infty\\ \overset{②}{\leq}&\frac{1.0175}{\tau}\frac{0.00682\sqrt{\tau}}{1-0.00755}\leq\frac{0.00700}{\sqrt{\tau}},\end{aligned}\tag{A.63}$$

where ① follows from sub-multiplicative property of the operator norm $\|\cdot\|_{\infty,\infty}$, and ② follows from Eq. (A.61) and $\|\mathbf{s}^\star\|_\infty=1$. This indicates that

$$\|\boldsymbol{\beta}\|_\infty\leq\|\mathrm{diag}(\mathbf{s}^\star)\|_{\infty,\infty}\|\tilde{\boldsymbol{\beta}}\|_\infty\leq 0.00700/\sqrt{\tau}\leq 0.00386/n:=\beta^\infty,\tag{A.64}$$

where the last inequality follows because $\tau\geq 3.289n^2$ for $n\geq 130$ by (A.2).

A.6.2.2 Bounding $\|\boldsymbol{\alpha}\|_\infty$ and $\mathbb{R}\{\alpha_1\}$ and $|\mathbb{I}\{\alpha_1\}|$

From (A.58), we have

$$\begin{aligned}\|\boldsymbol{\alpha}\|_\infty\overset{①}{\leq}&\|\mathbf{D}_0(\mathbf{f}^\star)^{-1}\|_{\infty,\infty}\|\mathbf{s}^\star\|_\infty+\|\mathbf{D}_0(\mathbf{f}^\star)^{-1}\|_{\infty,\infty}\|\mathbf{D}_1(\mathbf{f}^\star)\|_{\infty,\infty}\|\boldsymbol{\beta}\|_\infty\\ \overset{②}{\leq}&\frac{1}{1-0.00755}+\frac{0.00682\sqrt{\tau}}{1-0.00755}\frac{0.00700}{\sqrt{\tau}}\\ \leq&1.00766:=\alpha^\infty,\end{aligned}\tag{A.65}$$

where ① follows from the triangle inequality and the fact that $\|\mathbf{ABx}\|_\infty\leq\|\mathbf{A}\|_{\infty,\infty}\|\mathbf{B}\|_{\infty,\infty}\|\mathbf{x}\|_\infty$. ② holds since $\|\mathbf{s}^\star\|_\infty=1$.

Second, recognizing that $\alpha_1=1-[(\mathbf{I}-\mathbf{D}_0(\mathbf{f}^\star)^{-1})\mathbf{s}^\star-\mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{D}_1(\mathbf{f}^\star)^H\boldsymbol{\beta}]_1$ by Eq. (A.59), we have $\mathbb{R}\{\alpha_1\}=1-[\mathbb{R}\{(\mathbf{I}-\mathbf{D}_0(\mathbf{f}^\star)^{-1})\mathbf{s}^\star-\mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{D}_1(\mathbf{f}^\star)^H\boldsymbol{\beta}\}]_1$. We further get an upper bound as follows

$$
\begin{aligned}
&\left|\left[\mathbb{R}\left\{(\mathbf{I}-\mathbf{D}_0(\mathbf{f}^\star)^{-1})\mathbf{s}^\star-\mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{D}_1(\mathbf{f}^\star)^H\boldsymbol{\beta}\right\}\right]_1\right| \\
\overset{①}{\leq}&\|(\mathbf{I}-\mathbf{D}_0(\mathbf{f}^\star)^{-1})\mathbf{s}^\star-\mathbf{D}_0(\mathbf{f}^\star)^{-1}\mathbf{D}_1(\mathbf{f}^\star)^H\boldsymbol{\beta}\|_\infty \\
\overset{②}{\leq}&\|\mathbf{D}_0(\mathbf{f}^\star)^{-1}\|_{\infty,\infty}\|\mathbf{I}-\mathbf{D}_0(\mathbf{f}^\star)\|_{\infty,\infty}\|\mathbf{s}^\star\|_\infty+\|\mathbf{D}_0(\mathbf{f}^\star)^{-1}\|_{\infty,\infty}\|\mathbf{D}_1(\mathbf{f}^\star)\|_{\infty,\infty}\|\boldsymbol{\beta}\|_\infty \\
\leq&\frac{0.00755}{1-0.00755}+\frac{0.00682\sqrt{\tau}}{1-0.00755}\frac{0.00700}{\sqrt{\tau}} \\
\leq&0.00766,
\end{aligned}
$$

where ① follows from the real part of the first entry of a vector is no larger than the infinity norm of this vector and ② follows from the triangle inequality and the sub-multiplicative property of infinity operator norm that $\|\mathbf{ABx}\|_\infty \leq \|\mathbf{A}\|_{\infty,\infty}\|\mathbf{B}\|_{\infty,\infty}\|\mathbf{x}\|_\infty$. The last inequality follows from Eq. (A.61) and (A.64). Combining the above arguments yields

$$
\mathbb{R}\{\alpha_1\} \geq 1-0.00766 \quad \text{and} \quad |\mathbb{I}\{\alpha_1\}| \leq 0.00766. \tag{A.66}
$$

We are ready to show the Boundedness property following the simplifications used in [13]. In particular, fix an arbitrary point $f_0^\star \in T^\star$ as the reference point and let $f_{-1}^\star$ be the first frequency in $T^\star$ that lies on the left of $f_0^\star$ while $f_1^\star$ be the first frequency in $T^\star$ that lies on the right. Here "left" and "right" are directions on the complex circle $\mathbb{T}$. We remark that the analysis depends only on the relative locations of $\{f_\ell^\star\}$. Hence, to simplify the arguments, we assume that the reference point $f_0^\star$ is at 0 by shifting the frequencies if necessary. Then we divide the region between $f_0^\star = 0$ and $f_1^\star/2$ into three parts: Near Region $\mathcal{N} := [0, 0.24/n]$, Middle Region $\mathcal{M} := [0.24/n, 0.75/n]$ and Far Region $\mathcal{F} := [0.75/n, f_1^\star/2]$. Also their symmetric counterparts: $-\mathcal{N} := [-0.24/n, 0]$, $-\mathcal{M} := [-0.75/n, -0.24/n]$, and $-\mathcal{F} := [f_{-1}^\star/2, -0.75/n]$. We first show that the dual polynomial has strictly negative curvature $|Q^\star(f)|'' < 0$ in $\mathcal{N} = [0, 0.24/n]$ and $|Q^\star(f)| < 1$ in $\mathcal{M} \cup \mathcal{F} = [0.24/n, f_1^\star/2]$, implying $|Q^\star(f)| < 1$ in $\mathcal{N} \cup \mathcal{M} \cup \mathcal{F} \backslash \{f_0^\star\}$ by exploiting $|Q^\star(f_0^\star)| = 1$ and $|Q^\star(f_0^\star)|' = 0$. Then using the same symmetric arguments in [13], we claim that $|Q^\star(f)| < 1$ in $(-\mathcal{N}) \cup (-\mathcal{M}) \cup (-\mathcal{F}) \backslash \{f_0^\star\}$. Combining these two results with the fact that the reference point $f_0^\star$ is chosen arbitrarily from $T^\star$ (and shifted to 0), we establish that the Boundedness property of $Q^\star(f)$ holds in the entire $\mathbb{T} \backslash T^\star$.

A.6.2.3 Controlling $Q^\star(f)$ in Near Region

For $f \in \mathcal{N}$, the second-order Taylor expansion of $|Q^\star(f)|$ at $f_0^\star = 0$ states

$$
\begin{aligned}
|Q^\star(f)| &= |Q^\star(f_0^\star)| + (f - f_0^\star)|Q^\star(f_0^\star)|' + \frac{1}{2}(f - f_0^\star)^2|Q^\star(\xi)|'' \\
&= 1 + (f - f_0^\star)|Q^\star(f_0^\star)|' + \frac{1}{2}(f - f_0^\star)^2|Q^\star(\xi)|'' \text{ for some } \xi \in \mathcal{N},
\end{aligned} \tag{A.67}
$$

with the second line following from the Interpolation property. We argue that

$$|Q^\star(f_0^\star)|' = \frac{Q_R^\star(f_0^\star)Q_R^\star(f_0^\star)' + Q_I^\star(f_0^\star)Q_I^\star(f_0^\star)'}{|Q^\star(f_0^\star)|} = \frac{\mathbb{R}\{c_0^\star\}Q_R^\star(f_0^\star)' + \mathbb{I}\{c_0^\star\}Q_I^\star(f_0^\star)'}{|c_0^\star||Q^\star(f_0^\star)|} = \frac{\mathbb{R}\{c_0^{\star H}Q^\star(f_0^\star)'\}}{|c_0^\star||Q^\star(f_0^\star)|} = 0.$$

The last equality is due to (A.55). Together with (A.67), to bound $|Q^\star(f)|$ strictly below 1, we only need to show the concavity of $|Q^\star(f)|$ in Near Region (i.e., $|Q^\star(f)|'' < 0$ for $f \in \mathcal{N}$). Since

$$|Q^\star(f)|'' = -\frac{(Q_R^\star(f)Q_R^\star(f)' + Q_I^\star(f)Q_I^\star(f)')^2}{|Q^\star(f)|^3} + \frac{Q_R^\star(f)Q_R^\star(f)'' + |Q^\star(f)'|^2 + |Q_I^\star(f)||Q_I^\star(f)''|}{|Q^\star(f)|},$$

we only need to show that

$$Q_R^\star(f)Q_R^\star(f)'' + |Q^\star(f)'|^2 + |Q_I^\star(f)||Q_I^\star(f)''| < 0.$$

Recall the expression for $Q^\star(f)$ given in Eq. (A.53)

$$Q^\star(f) = \sum_{f_\ell^\star \in T^\star} \alpha_\ell K(f_\ell^\star - f) + \sum_{f_\ell^\star \in T^\star} \beta_\ell K'(f_\ell^\star - f).$$

To bound the real part of $Q^\star(f)$ in $\mathcal{N} = [0, 0.24/n]$, we observe

$$\begin{aligned}
Q_R^\star(f) &\geq \mathbb{R}\{\alpha_1 K(f)\} - \alpha^\infty \sum_{f_\ell^\star \in T^\star \setminus \{0\}} |K(f - f_\ell^\star)| - \beta^\infty |K'(f)| - \beta^\infty \sum_{f_\ell^\star \in T^\star \setminus \{0\}} |K'(f - f_\ell^\star)| \\
&\geq \mathbb{R}\{\alpha_1\} \min_{f \in \mathcal{N}} K(f) - \alpha^\infty F_0(2.5/n, f) - \beta^\infty (\max_{f \in \mathcal{N}} |K'(f)| + F_1(2.5/n, f)) \\
&\geq (1 - 0.00766)(0.905252) - (1.00766)0.00757 - (0.00386/n)(0.789569n + 0.01241n) \\
&\geq 0.887594,
\end{aligned}$$

where the first inequality follows from an application of the triangle inequality, and the second is from Lemma A.1.2. The third inequality follows from evaluating $F_0(2.5/n, f)$ and $F_1(2.5/n, f)$ at $f = 0.24/n$, the numerical bounds in Table A.1 and Table A.3 of Appendix A.1.5 and Eq. (A.64), (A.65), (A.66), as well as $\min_{f \in \mathcal{N}} K(f) \geq 0.905252$. This last bound follows from [13, Eq. (2.20), set $f_c = n - 2$] that $K(f) \geq 1 - \frac{\pi^2}{6}(n-2)(n+2)f^2$. Hence

$$\min_{f \in \mathcal{N}} K(f) \geq \min_{f \in \mathcal{N}} 1 - \frac{\pi^2}{6}(n-2)(n+2)f^2 \geq 1 - \frac{\pi^2}{6}(n-2)(n+2)(0.24/n)^2 \geq 0.905252.$$

Similarly, combining Eq. (A.64), (A.65), (A.66), the upper bounds on $F_\ell(2.5/n, 0.24/n)$ in Table A.1 and the upper bounds for $\max_{f \in \mathcal{N}} |K^{(\ell)}(f)|$ and $\max_{f \in \mathcal{N}} K''(f)$ in Table A.3, we get

$$\begin{aligned}
Q_R^\star{}''(f) \leq& \mathbb{R}\{\alpha_1\} \max_{f \in \mathcal{N}} K''(f) + \alpha^\infty F_2(2.5/n, 0.24/n) + \beta^\infty (\max_{f \in \mathcal{N}} |K'''(f)| + F_3(2.5/n, 0.24/n)) \\
\leq& (1 - 0.00766)(-2.35084n^2) + (1.00766)(0.05637n^2) + (0.00386/n)(7.79273n^3 + 0.28838n^3) \\
\leq& -2.24483n^2;
\end{aligned}$$

$$
\begin{aligned}
|Q_I^\star(f)| \leq & |\mathbb{I}\{\alpha_1\}| \max_{f\in\mathcal{N}} K(f) + \alpha^\infty F_0(2.5/n, 0.24/n) + \beta^\infty(\max_{f\in\mathcal{N}} |K'(f)| + F_1(2.5/n, 0.24/n)) \\
\leq & (0.00766) \times 1 + (1.00766)0.00757 + (0.00386/n)(0.789569n + 0.01241n) \\
\leq & 0.0183836; \\
|Q_I^{\star\prime\prime}(f)| \leq & |\mathbb{I}\{\alpha_1\}| \max_{f\in\mathcal{N}} |K''(f)| + \alpha^\infty F_2(2.5/n, 0.24/n) + \beta^\infty(\max_{f\in\mathcal{N}} |K'''(f)| + F_3(2.5/n, 0.24/n)) \\
\leq & (0.00766)(3.290n^2) + (1.00766)(0.05637n^2) + (0.00386/n)(7.79273n^3 + 0.28838n^3) \\
\leq & 0.113197n^2; \\
|Q^{\star\prime}(f)| \leq & \alpha^\infty(\max_{f\in\mathcal{N}} |K'(f)| + F_1(2.5/n, 0.24/n)) + \beta^\infty(\max_{f\in\mathcal{N}} |K''(f)| + F_2(2.5/n, 0.24/n)) \\
\leq & (1.00766)(0.789569n + 0.01241n) + (0.00386/n)(3.290n^2 + 0.05637n^2) \\
\leq & 0.821039n; \\
|Q^{\star\prime\prime}(f)| \leq & \alpha^\infty(\max_{f\in\mathcal{N}} |K''(f)| + F_2(2.5/n, 0.24/n)) + \beta^\infty(\max_{f\in\mathcal{N}} |K'''(f)| + F_3(2.5/n, 0.24/n)) \\
\leq & (1.00766)(3.290n^2 + 0.05637n^2) + (0.00386/n)(7.79273n^3 + 0.28838n^3) \\
\leq & 3.40320n^2.
\end{aligned}
$$

Combining the lower bound on $Q_R^\star(f)$ and the upper bounds on $Q_R^\star(f)''$, $|Q^\star(f)'|$, $|Q_I^\star(f)|$ and $|Q_I^\star(f)''|$, we arrive at

$$
|Q^\star(f)|'' = Q_R^\star(f)Q_R^\star(f)'' + |Q^\star(f)'|^2 + |Q_I^\star(f)||Q_I^\star(f)''| \leq -1.316313n^2 < 0 \text{ in } \mathcal{N}.
$$

A.6.2.4 Bounding $|Q^\star(f)|$ in Middle Region

For upperbounding $|Q^\star(f)|$ for $f \in \mathcal{M} = [0.24/n, 0.75/n]$, we firstly apply the triangle inequality

$$
\begin{aligned}
|Q^\star(f)| = & |\sum_{f_\ell^\star \in T^\star} \alpha_\ell K(f_\ell^\star - f) + \sum_{f_\ell^\star \in T^\star} \beta_\ell K'(f_\ell^\star - f)| \\
\leq & \|\boldsymbol{\alpha}\|_\infty \Big(|K(f)| + \sum_{f_\ell^\star \in T^\star \setminus \{0\}} |K(f - f_\ell^\star)| \Big) + \|\boldsymbol{\beta}\|_\infty \Big(|K'(f)| + \sum_{f_\ell^\star \in T^\star \setminus \{0\}} |K'(f - f_\ell^\star)| \Big) \\
\leq & \alpha^\infty |K(f)| + \beta^\infty |K'(f)| + \alpha^\infty F_0(2.5/n, f) + \beta^\infty F_1(2.5/n, f),
\end{aligned} \tag{A.68}
$$

where the last inequality is from Lemma A.1.2. We then follow [13, Eq. (2.29)] to upperbound the first two terms in the last line

$$
|K(f)| \leq 1 - \frac{\pi^2(n^2-4)f^2}{6} + \frac{\pi^4 n^4 f^4}{72} \qquad \text{and} \qquad |K'(f)| \leq \frac{\pi^2(n^2-4)f}{3}, \qquad \text{for } f \in [-1/2, 1/2].
$$

The rest of argument consists of defining

$$L_1(f) = \alpha^\infty \left(1 - \frac{1}{6}\pi^2(n^2-4)f^2 + \frac{1}{72}\pi^4 n^4 f^4\right) + \beta^\infty \frac{1}{3}\pi^2(n^2-4)f;$$
$$L_2(f) = \alpha^\infty F_0(2.5/n, f) + \beta^\infty F_1(2.5/n, f)$$

with the derivative of $L_1(f)$ given by

$$L_1'(f) = -\alpha^\infty \left(\frac{\pi^2(n^2-4)f}{3} - \frac{\pi^4 n^4 f^3}{18}\right) + \beta^\infty \frac{\pi^2(n^2-4)}{3} < 0, \quad \text{for } f \in \mathcal{M},$$

implying that $L_1(f)$ is decreasing. Also, $L_2(f)$ is increasing in $\mathcal{M}$ by Lemma A.1.2. Hence, by the monotonic property, we have

$$|Q^\star(f)| \leq L_1(0.24/n) + L_2(0.75/n) \leq 0.919779 + 0.007836 = 0.927615 < 1.$$

Bounding $|Q^\star(f)|$ in Far Region.

Recall that $f_0^\star = 0$ is the reference point. To simplify notation, we re-index the frequencies such that $\ldots \leq f_{-1}^\star < f_0^\star = 0 < f_1^\star < \ldots$. For $f \in \mathcal{F} = [0.75/n, f_1^\star/2] = [0.75/n, f_1^\star - f_1^\star/2]$, by Lemma A.1.3, we have

$$\begin{aligned}
\sum_j |K^{(\ell)}(f - f_j^\star)| \leq W_\ell(0.75/n, f_1^\star/2) &\overset{①}{=} \sum_{j\geq 0} B_\ell(j(2.5/n) + 0.75/n) + \sum_{j\geq 0} B_\ell(j(2.5/n) + f_1^\star/2) \\
&\overset{②}{\leq} \sum_{j\geq 0} B_\ell(j(2.5/n) + 0.75/n) + \sum_{j\geq 0} B_\ell(j(2.5/n) + 1.25/n) \\
&\overset{③}{=} W_\ell(0.75/n, 1.25/n), \qquad \text{(A.69)}
\end{aligned}$$

where ① follow from the definition of $W(\underline{f}, \bar{f})$ in Lemma A.1.3, ② follows $f_1^\star/2 = (f_1^\star - f_0^\star)/2 \geq \Delta_{\min}/2 = 1.25/n$ and decreasing property of $B_\ell(\cdot)$, and ③ follows from the definition of $W(\underline{f}, \bar{f})$.

Finally, applying (A.69), (A.64) and (A.65) to (A.53), we arrive at

$$\begin{aligned}
|Q^\star(f)| \leq& \alpha^\infty \sum_\ell |K(f - f_\ell^\star)| + \beta^\infty \sum_\ell |K'(f - f_\ell^\star)| \\
\leq& 1.00766 W_0(0.75/n, 1.25/n) + (0.00386/n) W_1(0.75/n, 1.25/n) \\
\leq& 1.00766(0.70859) + (0.00386/n)(5.2084n) \\
=& 0.734123.
\end{aligned}$$

This concludes the proof of Lemma 2.4.3. □

A.7 Proof of Lemma 2.4.4

Lemma A.7.1 (Lemma 2.4.4)**.** *Under the settings of Lemma 2.4.1, let $Q^\lambda(f)$ and $Q^\star(f)$ be the dual polynomials corresponding to $\boldsymbol{\theta}^\lambda$ and $\boldsymbol{\theta}^\star$, respectively. Then the distances between $Q^\lambda(f)$ and $Q^\star(f)$ and their various derivatives*

are uniformly bounded:

$$
\begin{array}{ll}
|Q^\star(f) - Q^\lambda(f)| \le 28.7343 X^\star B^\star \gamma,\ f \in \mathcal{N}, & |Q^\star(f) - Q^\lambda(f)| \le 39.3557 X^\star B^\star \gamma,\ f \in \mathcal{M}, \\
|Q^{\star\prime}(f) - Q^{\lambda'}(f)| \le 44.4648 n X^\star B^\star \gamma,\ f \in \mathcal{N}, & |Q^\star(f) - Q^\lambda(f)| \le 66.1596 X^\star B^\star \gamma,\ f \in \mathcal{F}, \\
|Q^{\star\prime\prime}(f) - Q^{\lambda''}(f)| \le 140.808 n^2 X^\star B^\star \gamma,\ f \in \mathcal{N}. &
\end{array}
$$

Proof. We exploit the closeness of $\boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}^\lambda$ (see Lemma 2.4.1) to bound the pointwise distance between $Q^\star(f)$ and $Q^\lambda(f)$. Note

$$
Q^\lambda(f) - Q^\star(f) = \mathbf{a}(f)^H \mathbf{Z}(\mathbf{q}^\lambda - \mathbf{q}^\star) = \mathbf{a}(f)^H \mathbf{Z} \left(\frac{\mathbf{x}^\star - \mathbf{x}^\lambda}{\lambda} + \frac{\mathrm{d}}{\mathrm{d}\lambda} \mathbf{x}^\lambda \big|_{\lambda=0} \right) = \frac{1}{\lambda} \int_0^\lambda \mathbf{a}(f)^H \mathbf{Z} \left(\frac{\mathrm{d}}{\mathrm{d}t} \mathbf{x}^\star - \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{x}^t \right) \mathrm{d}t,
$$

which implies that

$$
|Q^\lambda(f) - Q^\star(f)| \le \max_{0 \le t \le \lambda} \left| \mathbf{a}(f)^H \mathbf{Z} (\frac{\mathrm{d}}{\mathrm{d}t} \mathbf{x}^\star - \mathbf{a}(f)^H \mathbf{Z} (\frac{\mathrm{d}}{\mathrm{d}t} \mathbf{x}^t \right|. \tag{A.70}
$$

We can also obtain similar bounds on the pointwise distances between derivatives of $Q^\lambda(f)$ and $Q^\star(f)$.

Recall from Eq. (A.47), (2.20), and (A.50) that

$$
\frac{\mathrm{d}}{\mathrm{d}\lambda} \mathbf{x}^\lambda = -[\mathbf{A}'(\mathbf{f}^\lambda) \operatorname{diag}(\mathbf{c}^\lambda) \quad \mathbf{A}(\mathbf{f}^\lambda) \quad i\mathbf{A}(\mathbf{f}^\lambda)] (\nabla^2 G^\lambda(\boldsymbol{\theta}^\lambda))^{-1} \boldsymbol{\rho}^\lambda, \tag{A.71}
$$

$$
\frac{\mathrm{d}}{\mathrm{d}\lambda} \mathbf{x}^\star = -[\mathbf{A}'(\mathbf{f}^\star) \operatorname{diag}(\mathbf{c}^\star) \quad \mathbf{A}(\mathbf{f}^\star) \quad i\mathbf{A}(\mathbf{f}^\star)] (\nabla^2 G^0(\boldsymbol{\theta}^\star))^{-1} \boldsymbol{\rho}^\star, \tag{A.72}
$$

where $\boldsymbol{\rho}^\star = \begin{bmatrix} \mathbf{0}^T & \mathbb{R}\{\operatorname{sign}(\mathbf{c}^\star)\}^T & \mathbb{I}\{\operatorname{sign}(\mathbf{c}^\star)\}^T \end{bmatrix}^T$ and $\boldsymbol{\rho}^\lambda = \begin{bmatrix} \mathbf{0}^T & \mathbb{R}\{\operatorname{sign}(\mathbf{c}^\lambda)\}^T & \mathbb{I}\{\operatorname{sign}(\mathbf{c}^\lambda)\}^T \end{bmatrix}^T$.

Multiplying both sides of Eq. (A.71) and (A.72) by $-\mathbf{a}(f)^H \mathbf{Z}($ and then inserting $\mathbf{W}^{\star\frac{1}{2}} \mathbf{W}^{\star-\frac{1}{2}}$ (which equals $\mathbf{I}$) into the spaces before and after $(\nabla^2 G^0(\boldsymbol{\theta}^\star))^{-1}$ (and $(\nabla^2 G^\lambda(\boldsymbol{\theta}^\star))^{-1}$) yield

$$
\begin{aligned}
-\mathbf{a}(f)^H \mathbf{Z} (\frac{\mathrm{d}}{\mathrm{d}\lambda} \mathbf{x}^\lambda &= \boldsymbol{\nu}^\lambda(f) \Xi^\lambda \boldsymbol{\rho}^\lambda, \\
-\mathbf{a}(f)^H \mathbf{Z} (\frac{\mathrm{d}}{\mathrm{d}\lambda} \mathbf{x}^\star &= \boldsymbol{\nu}^\star(f) \Xi^\star \boldsymbol{\rho}^\star.
\end{aligned}
$$

Here

$$
\begin{aligned}
\boldsymbol{\nu}^\lambda(f) &:= [\mathbf{D}_1(f, \mathbf{f}^\lambda) \operatorname{diag}(\mathbf{c}^\lambda) \mathbf{S}^{-1} \ \ \mathbf{D}_0(f, \mathbf{f}^\lambda) \ i\mathbf{D}_0(f, \mathbf{f}^\lambda)], \\
\boldsymbol{\nu}^\star(f) &:= [\mathbf{D}_1(f, \mathbf{f}^\star) \operatorname{diag}(\mathbf{c}^\star) \mathbf{S}^{-1} \ \ \mathbf{D}_0(f, \mathbf{f}^\star) \ i\mathbf{D}_0(f, \mathbf{f}^\star)],
\end{aligned} \tag{A.73}
$$

with $\mathbf{D}_\ell(f, \mathbf{f}^\lambda)$ a row vector defined by $\mathbf{D}_\ell(f, \mathbf{f}^\lambda) := [K_\ell(f_1^\lambda - f), \ldots, K_\ell(f_k^\lambda - f)]$, and

$$
\begin{aligned}
\Xi^\lambda &:= \Upsilon (\nabla^2 G^\lambda(\boldsymbol{\theta}^\lambda))^{-1}, \\
\Xi^\star &:= \Upsilon (\nabla^2 G^0(\boldsymbol{\theta}^\star))^{-1},
\end{aligned}
$$

where $\Upsilon(\cdot) := \mathbf{W}^{\star\frac{1}{2}}(\cdot)\mathbf{W}^{\star\frac{1}{2}}$ is a linear operator that normalizes the Hessian matrix so that it is close to the identity. As a consequence, we bound the integrand of (A.70) as follows

$$\begin{aligned}
&|\boldsymbol{\nu}^\star(f)\Xi^\star\boldsymbol{\rho}^\star - \boldsymbol{\nu}^\lambda(f)\Xi^\lambda\boldsymbol{\rho}^\lambda| \\
\leq&|\boldsymbol{\nu}^\star(f)\Xi^\star(\boldsymbol{\rho}^\star - \boldsymbol{\rho}^\lambda)| + |\boldsymbol{\nu}^\star(f)(\Xi^\star - \Xi^\lambda)\boldsymbol{\rho}^\lambda| + |(\boldsymbol{\nu}^\star(f) - \boldsymbol{\nu}^\lambda(f))\Xi^\lambda\boldsymbol{\rho}^\lambda| \\
\leq&\|\boldsymbol{\nu}^\star(f)\|_1\|\Xi^\star\|_{\infty,\infty}\|\boldsymbol{\rho}^\star - \boldsymbol{\rho}^\lambda\|_\infty + \|\boldsymbol{\nu}^\star(f)\|_1\|\Xi^\star - \Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty + \|\boldsymbol{\nu}^\star(f) - \boldsymbol{\nu}^\lambda(f)\|_1\|\Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty,
\end{aligned} \tag{A.74}$$

where the first line follows from the triangle inequality and the second line follows from Hölder's inequality and the sub-multiplicative property of the $\ell_{\infty,\infty}$ norm. We next develop upper bounds on $\|\boldsymbol{\nu}^\star(f)\|_1$, $\|\boldsymbol{\nu}^\star(f) - \boldsymbol{\nu}^\lambda(f)\|_1$, $\|\Xi^\star\|_{\infty,\infty}$, $\|\Xi^\star - \Xi^\lambda\|_{\infty,\infty}$, $\|\Xi^\lambda\|_{\infty,\infty}$, $\|\boldsymbol{\rho}^\star - \boldsymbol{\rho}^\lambda\|_1$ and $\|\boldsymbol{\rho}^\lambda\|_\infty$.

Bounding $\|\Xi^\star\|_{\infty,\infty}$ and $\|\Xi^\lambda\|_{\infty,\infty}$ and $\|\Xi^\star - \Xi^\lambda\|_{\infty,\infty}$.

We note that both $\Xi^{\star^{-1}} = \Upsilon(\nabla^2 G^0(\boldsymbol{\theta}^\star)))$ and $\Xi^{\lambda^{-1}} = \Upsilon(\nabla^2 G^\lambda(\boldsymbol{\theta}^\lambda)))$ are close to the identity matrix. More precisely, we have

$$\begin{aligned}
\|\mathbf{I} - \Xi^{\star^{-1}}\|_\infty \overset{①}{=}& \|\mathbf{I} - \Upsilon(\nabla^2 G^0(\boldsymbol{\theta}^\star))\|_{\infty,\infty} \\
\overset{②}{\leq}& [\|\mathbf{I} - \operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)^H(-\mathbf{D}_2(\mathbf{f}^\star)/\tau)\operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)\|_{\infty,\infty} + 2\|\operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)\mathbf{D}_1(\mathbf{f}^\star)/\sqrt{\tau}\|_{\infty,\infty}] \\
&\vee (\|\operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)\mathbf{D}_1(\mathbf{f}^\star)/\sqrt{\tau}\|_{\infty,\infty} + \|\mathbf{I} - \mathbf{D}_0(\mathbf{f}^\star)\|_{\infty,\infty}) \\
\overset{③}{\leq}& \|\mathbf{I} - (-\mathbf{D}_2(\mathbf{f}^\star)/\tau)\|_{\infty,\infty} + 2\|\mathbf{D}_1(\mathbf{f}^\star)/\sqrt{\tau}\|_{\infty,\infty} \\
\overset{④}{\leq}& 0.0171 + 2\times 0.00682 \\
\leq& 0.03074,
\end{aligned}$$

where $a \vee b := \max(a, b)$. ① follows from definition of $\Xi^\star$ and ② follows from applying the triangle inequality to the expression of $[\mathbf{I} - \Upsilon(\nabla^2 G^0(\boldsymbol{\theta}^\star))]$. ③ follows since the infinity norm of any sign vector is 1, bounding $\|\operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)\,\mathbf{D}_1(\mathbf{f}^\star)/\sqrt{\tau}\|_{\infty,\infty}$ is equivalent to bound $\|\mathbf{D}_1(\mathbf{f}^\star)/\sqrt{\tau}\|_{\infty,\infty}$. Finally, ④ follows from Eq. (A.11). This leads to

$$\|\Xi^\star\|_{\infty,\infty} \leq \frac{1}{1 - \|\mathbf{I} - \Xi^{\star^{-1}}\|_{\infty,\infty}} \leq \frac{1}{1 - 0.03074} \leq 1.03172. \tag{A.75}$$

According to (A.41), we have

$$\|\mathbf{I} - \Xi^{\lambda^{-1}}\|_{\infty,\infty} = \|I - \Upsilon(\nabla^2 G^\lambda(\boldsymbol{\theta}^\lambda))\|_{\infty,\infty} \leq 0.08561,$$

yielding

$$\|\Xi^\lambda\|_{\infty,\infty} \leq \frac{1}{1 - \|\mathbf{I} - \Xi^{\lambda^{-1}}\|_{\infty,\infty}} \leq \frac{1}{1 - 0.08561} \leq 1.09363. \tag{A.76}$$

Next, note

$$\|\Xi^{\star^{-1}} - \Xi^{\lambda^{-1}}\|_{\infty,\infty} = \|\Upsilon(\nabla^2 G^0(\boldsymbol{\theta}^\star)) - \Upsilon(\nabla^2 G^\lambda(\boldsymbol{\theta}^\lambda))\|_{\infty,\infty} \leq \max\{\Pi_1, \Pi_2, \Pi_3\},$$

where Π_1, Π_2, Π_3 denote the first, second and third absolute row block sums of $[\Upsilon(\nabla^2 G^0(\boldsymbol{\theta}^\star)) - \Upsilon(\nabla^2 G^\lambda(\boldsymbol{\theta}^\lambda))]$. We first bound Π_1 as follows

$$\begin{aligned}
\Pi_1 =& \| \operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|)^H \mathbf{D}_2(\mathbf{f}) \operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|) - \operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)^H \mathbf{D}_2(\mathbf{f}^\star) \operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)\|_{\infty,\infty}/\tau \\
&+ 2 \left\|\operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|)^H \mathbf{D}_1(\mathbf{f}) - \operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)^H \mathbf{D}_1(\mathbf{f}^\star)\right\|_{\infty,\infty}/\sqrt{\tau} \\
&+ 2 \|\operatorname{diag}(1./|\mathbf{c}^\star|)\left[\mathbf{D}_1(\mathbf{f}, \mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f})\mathbf{c}\right]\|_\infty/\sqrt{\tau} \\
&+ \| \operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|^2)^H [\mathbf{D}_2(\mathbf{f}, \mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_2(\mathbf{f})\mathbf{c}]\|_\infty/\tau \\
\overset{①}{\leq}& [2.19778 X^\star\gamma + 1.14168(X^\star\gamma)^2] + 2[1.48286 X^\star\gamma + 1.47604(X^\star\gamma)^2] + 2(0.75038)X^\star B^\star\gamma + 1.14168 X^\star B^\star\gamma \\
\leq& 7.81004 X^\star B^\star\gamma \text{ (by } B^\star X^\star\gamma \leq 10^{-3}\text{)}, \qquad \text{(A.77)}
\end{aligned}$$

where ① follows from combining Eq. (A.19)- (A.20) and (A.78)- (A.79), where (A.78)- (A.79) are given by

$$\begin{aligned}
&\| \operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|)^H \mathbf{D}_2(\mathbf{f}) \operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|) - \operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)^H \mathbf{D}_2(\mathbf{f}^\star) \operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)\|_{\infty,\infty}/\tau \\
\leq& \| \operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|)^H \mathbf{D}_2(\mathbf{f}) \operatorname{diag}((\mathbf{c} - \mathbf{c}^\star)./|\mathbf{c}^\star|)\|_{\infty,\infty}/\tau \\
&+ \| \operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|)^H (\mathbf{D}_2(\mathbf{f}) - \mathbf{D}_2(\mathbf{f}^\star)) \operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)\|_{\infty,\infty}/\tau \\
&+ \| \operatorname{diag}((\mathbf{c} - \mathbf{c}^\star)./|\mathbf{c}^\star|)^H \mathbf{D}_2(\mathbf{f}^\star) \operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)\|_{\infty,\infty}/\tau \\
\leq& (1 + X^\star\gamma)(1.05610)(X^\star\gamma) + (1 + X^\star\gamma)(0.08558 X^\star\gamma) + (X^\star\gamma)(1.05610) \qquad \text{(by (A.18) and (A.11))} \\
\leq& 2.19778 X^\star\gamma + 1.14168(X^\star\gamma)^2 \qquad \text{(A.78)}
\end{aligned}$$

and

$$\begin{aligned}
&\left\|\operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|)^H \mathbf{D}_1(\mathbf{f}) - \operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)^H \mathbf{D}_1(\mathbf{f}^\star)\right\|_{\infty,\infty}/\sqrt{\tau} \\
\leq& \left\|\operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|)^H (\mathbf{D}_1(\mathbf{f}) - \mathbf{D}_1(\mathbf{f}^\star))\right\|_{\infty,\infty}/\sqrt{\tau} + \left\|\operatorname{diag}((\mathbf{c} - \mathbf{c}^\star)./|\mathbf{c}^\star|)^H \mathbf{D}_1(\mathbf{f}^\star)\right\|_{\infty,\infty}/\sqrt{\tau} \\
\leq& (1 + X^\star\gamma)(1.47604 X^\star\gamma) + (X^\star\gamma)(0.00682) \qquad \text{(by Eq. (A.17) and (A.11))} \\
\leq& 1.48286 X^\star\gamma + 1.47604(X^\star\gamma)^2. \qquad \text{(A.79)}
\end{aligned}$$

We next bound Π_2 and Π_3:

$$\begin{aligned}
\{\Pi_2, \Pi_3\} \overset{①}{\leq}& \|\mathbf{D}_1(\mathbf{f}) \operatorname{diag}(\mathbf{c}./|\mathbf{c}^\star|) - \mathbf{D}_1(\mathbf{f}^\star) \operatorname{diag}(\mathbf{c}^\star./|\mathbf{c}^\star|)\|_{\infty,\infty}/\sqrt{\tau} \\
&+ \|\mathbf{D}_0(\mathbf{f}) - \mathbf{D}_0(\mathbf{f}^\star)\|_{\infty,\infty} + \| \operatorname{diag}(1./|\mathbf{c}^\star|)[\mathbf{D}_1(\mathbf{f}, \mathbf{f}^\star)\mathbf{c}^\star - \mathbf{D}_1(\mathbf{f})\mathbf{c}]\|_\infty/\sqrt{\tau} \\
&+ \lambda\|\mathbf{u} \odot \mathbf{u}./|\mathbf{c}|^3 - \mathbf{u}^\star \odot \mathbf{u}^\star./|\mathbf{c}^\star|^3\|_\infty + \lambda\|\mathbf{u} \odot \mathbf{v}./|\mathbf{c}|^3 - \mathbf{u}^\star \odot \mathbf{v}^\star./|\mathbf{c}^\star|^3\|_\infty \\
\overset{②}{\leq}& [1.48286 X^\star\gamma + 1.47604(X^\star\gamma)^2] + 0.01516 X^\star\gamma + 0.75038 X^\star B^\star\gamma + 2(0.646 X^\star\gamma)(5.00701) X^\star\gamma \\
\leq& 2.25636 X^\star B\gamma \\
<& \Pi_1 \text{ (by } B^\star X^\star\gamma \leq 10^{-3}\text{)},
\end{aligned}$$

where ① follows from the triangle inequality and ② follows by combining Eq. (A.79), (A.16), (A.19), and (A.80). To show (A.80), we assume the norm $\|\mathbf{u}\odot\mathbf{u}./|\mathbf{c}|^3-\mathbf{u}^\star\odot\mathbf{u}^\star./|\mathbf{c}^\star|^3\|_\infty$ is achieved by the ℓth entry and proceed as

$$\begin{aligned}\left|\frac{u_\ell^2}{|c_\ell|^3}-\frac{u_\ell^{\star 2}}{|c_\ell^\star|^3}\right| &\overset{①}{\leq}\left|\frac{c_\ell^2}{|c_\ell|^3}-\frac{c_\ell^{\star 2}}{|c_\ell^\star|^3}\right| \\ &\overset{②}{\leq}\frac{|c_\ell^2-c_\ell^{\star 2}|}{|c_\ell^\star|^3}+|c_\ell|^2\left|\frac{1}{|c_\ell|^3}-\frac{1}{|c_\ell^\star|^3}\right| \\ &\overset{③}{\leq}\frac{X^\star\gamma}{c_{\min}^\star}\left((2+X^\star\gamma)+\frac{(X^\star\gamma)^2+3(X^\star\gamma)+3}{1-X^\star\gamma}\right)\leq\frac{X^\star\gamma}{c_{\min}^\star}(5.00701),\end{aligned} \tag{A.80}$$

where ① follows from $|\mathbb{R}\{a\}|\leq|a|$ for all $a\in\mathbb{C}$ and ② follows from the triangle inequality. ③ follows from Eq. (A.81) and (A.82) given below:

$$\frac{|c_\ell^2-c_\ell^{\star 2}|}{|c_\ell^\star|^3}\leq\frac{1}{|c_\ell^\star|}\frac{|c_\ell-c_\ell^\star|}{|c_\ell^\star|}\frac{|c_\ell+c_\ell^\star|}{|c_\ell^\star|}\leq\frac{X^\star\gamma(2+X^\star\gamma)}{c_{\min}^\star} \tag{A.81}$$

and

$$\begin{aligned}|c_\ell|^2\left|\frac{1}{|c_\ell|^3}-\frac{1}{|c_\ell^\star|^3}\right|=|c_\ell|^2\left|\frac{1}{|c_\ell|}-\frac{1}{|c_\ell^\star|}\right|\left(\frac{1}{|c_\ell|^2}+\frac{1}{|c_\ell^\star|^2}+\frac{1}{|c_\ell^\star||c_\ell|}\right)&\leq\frac{|c_\ell-c_\ell^\star|}{|c_\ell||c_\ell^\star|}\left(1+\frac{|c_\ell|^2}{|c_\ell^\star|^2}+\frac{|c_\ell|}{|c_\ell^\star|}\right) \\ &\leq\frac{1}{|c_\ell|}X^\star\gamma\left(1+\frac{|c_\ell|^2}{|c_\ell^\star|^2}+\frac{|c_\ell|}{|c_\ell^\star|}\right) \\ &\leq\frac{1}{c_{\min}^\star(1-X^\star\gamma)}X^\star\gamma\left(1+\frac{|c_\ell|^2}{|c_\ell^\star|^2}+\frac{|c_\ell|}{|c_\ell^\star|}\right) \\ &\leq\frac{X^\star\gamma}{c_{\min}^\star}\frac{(X^\star\gamma)^2+3(X^\star\gamma)+3}{1-X^\star\gamma}\end{aligned} \tag{A.82}$$

where the first line follows from $|a^3-b^3|=|(a-b)(a^2+ab+b^2)|=|a-b|(a^2+ab+b^2)$ for any positive a,b. The second line holds since $|\frac{1}{|c_\ell|}-\frac{1}{|c_\ell^\star|}|=\frac{||c_\ell|-|c_\ell^\star||}{|c_\ell||c_\ell^\star|}\leq\frac{|c_\ell-c_\ell^\star|}{|c_\ell||c_\ell^\star|}$ by the triangle inequality. The third line follows from $|c_\ell-c_\ell^\star|/|c_\ell^\star|\leq X^\star\gamma$ by (A.10). For the fourth line to hold, note that by (A.10), $\frac{|c_i-c_i^\star|}{|c_i^\star|}\leq X^\star\gamma$, which implies that $|c_i|\geq|c_i^\star|-|c_i-c_i^\star|\geq(1-X^\star\gamma|)c_{\min}^\star|$. The last line follows from $|c_\ell|/|c_\ell^\star|\leq(1+X^\star\gamma)$. Finally, we get the bound

$$\|\Xi^{\star-1}-\Xi^{\lambda^{-1}}\|_{\infty,\infty}=\Pi_1\leq 7.81004X^\star B^\star\gamma$$

implying

$$\begin{aligned}\|\Xi^\star-\Xi^\lambda\|_{\infty,\infty}&\leq\|\Xi^\star\|_{\infty,\infty}\|\Xi^{\star-1}-\Xi^{\lambda^{-1}}\|_{\infty,\infty}\|\Xi^\lambda\|_{\infty,\infty} \\ &\leq(1.03172)(1.09363)(7.81004X^\star B^\star\gamma)=8.81222X^\star B^\star\gamma.\end{aligned} \tag{A.83}$$

Bounding $\|\boldsymbol{\rho}^\star-\boldsymbol{\rho}^\lambda\|_\infty$ and $\|\boldsymbol{\rho}^\lambda\|_\infty$.

First recognize that $\|\boldsymbol{\rho}^\lambda\|_\infty=1$ since $\boldsymbol{\rho}^\lambda$ contains either signs or zeros. Assume the ℓ_∞ norm of $(\boldsymbol{\rho}^\star-\boldsymbol{\rho}^\lambda)$ is achieved

by $|\operatorname{sign}(c_\ell^\lambda) - \operatorname{sign}(c_\ell^\star)|$, then applying triangle inequalities gives

$$\begin{aligned}\|\boldsymbol{\rho}^\star - \boldsymbol{\rho}^\lambda\|_\infty = \left|\frac{c_\ell^\lambda}{|c_\ell^\lambda|} - \frac{c_\ell^\star}{|c_\ell^\star|}\right| = \left|\frac{c_\ell^\lambda}{|c_\ell^\lambda|} - \frac{c_\ell^\lambda}{|c_\ell^\star|} + \frac{c_\ell^\lambda}{|c_\ell^\star|} - \frac{c_\ell^\star}{|c_\ell^\star|}\right| \le& \left|\frac{c_\ell^\lambda}{|c_\ell^\lambda|} - \frac{c_\ell^\lambda}{|c_\ell^\star|}\right| + \frac{|c_\ell^\star - c_\ell^\lambda|}{|c_\ell^\star|} \\ =& |c_\ell^\lambda|\left|\frac{1}{|c_\ell^\lambda|} - \frac{1}{|c_\ell^\star|}\right| + \frac{|c_\ell^\star - c_\ell^\lambda|}{|c_\ell^\star|} \\ =& \left|c_\ell^\lambda\right|\left|\frac{|c_\ell^\lambda| - |c_\ell^\star|}{|c_\ell^\lambda||c_\ell^\star|}\right| + \frac{|c_\ell^\star - c_\ell^\lambda|}{|c_\ell^\star|} \\ \le& 2\frac{|c_\ell^\star - c_\ell^\lambda|}{|c_\ell^\star|} \le 2X^\star\gamma.\end{aligned} \tag{A.84}$$

Bounding $\boldsymbol{\nu}^\star(f), \boldsymbol{\nu}^\star(f)', \boldsymbol{\nu}^\star(f)''$ and $(\boldsymbol{\nu}^\star(f) - \boldsymbol{\nu}^\lambda(f)), (\boldsymbol{\nu}^\star(f) - \boldsymbol{\nu}^\lambda(f))', (\boldsymbol{\nu}^\star(f) - \boldsymbol{\nu}^\lambda(f))''$.
Applying the triangle inequality and the sub-multiplicative property of the norm to (A.74) and (A.73), we get

$$\begin{aligned}&\|\boldsymbol{\nu}^\star(f) - \boldsymbol{\nu}^\lambda(f)\|_1 \\ \le&\|[\mathbf{D}_1(f,\mathbf{f}^\lambda) - \mathbf{D}_1(f,\mathbf{f}^\star)]^T \operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1} - \mathbf{D}_1(f,\mathbf{f}^\lambda)^T\boldsymbol{\Phi}\|_1 + 2\|\mathbf{D}_0(f,\mathbf{f}^\lambda) - \mathbf{D}_0(f,\mathbf{f}^\star)\|_1 \\ \le&\|\mathbf{D}_1(f,\mathbf{f}^\lambda) - \mathbf{D}_1(f,\mathbf{f}^\star)\|_1\|\operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_{1,1} + \|\mathbf{D}_1(f,\mathbf{f}^\lambda)\|_1\|\boldsymbol{\Phi}\|_{1,1} + 2\|\mathbf{D}_0(f,\mathbf{f}^\lambda) - \mathbf{D}_0(f,\mathbf{f}^\star)\|_1; \\ &\|\boldsymbol{\nu}^\star(f)\|_1 \\ \le&\|\mathbf{D}_1(f,\mathbf{f}^\star)\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1}\|_1 + 2\|\mathbf{D}_0(f,\mathbf{f}^\star)\|_1 \le \|\mathbf{D}_1(f,\mathbf{f}^\star)\|_1\|\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1}\|_{1,1} + 2\|\mathbf{D}_0(f,\mathbf{f}^\star)\|_1,\end{aligned} \tag{A.85}$$

where $\boldsymbol{\Phi} := \operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1} - \operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}$ and $\mathbf{D}_\ell(f,\mathbf{f}) := [K^{(\ell)}(f_1 - f), \ldots, K^{(\ell)}(f_k - f)]$. Similar bounds also apply to various derivatives of $\boldsymbol{\nu}^\star(f)$ and $\boldsymbol{\nu}^\lambda(f)$, which we need in order to bound the distances between derivatives of $Q^\star(f)$ and $Q^\lambda(f)$. Using (A.10) and $\tau \ge 3.289n^2$, we have

$$\begin{aligned}\left\|(\operatorname{diag}(\mathbf{c}^\lambda) - \operatorname{diag}(\mathbf{c}^\star))\mathbf{S}^{-1}\right\|_{1,1} \le& (\max_i |c_i^\lambda - c_i^\star|/|c_i^\star|)/\sqrt{\tau} \le X^\star\gamma/\sqrt{\tau} \le 0.552X^\star\gamma/n; \\ \left\|\operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\right\|_{1,1} \le& (1 + X\gamma)/\sqrt{\tau} \le 0.552/n,\end{aligned} \tag{A.86}$$

which we need to continue the bounds in (A.85).

Since f may lie in different regions: Near Region, Middle Region, and Far Region, we next organize our analysis into three parts based on what region f is located in.

A.7.1 Near Region Analysis

We start with controlling $\|\mathbf{D}_\ell(f,\mathbf{f}^\lambda) - \mathbf{D}_\ell(f,\mathbf{f}^\star)\|_1$ and $|\mathbf{D}_\ell(f,\mathbf{f}^\star)\|_1$ for $\ell = 0,1,2,3$ in Near Region. When $\ell = 0$, we have

$$\begin{aligned}\|\mathbf{D}_0(f,\mathbf{f}^\lambda) - \mathbf{D}_0(f,\mathbf{f}^\star)\|_1 = \sum_\ell |K(f_\ell^\lambda - f) - K(f_\ell^\star - f)| \overset{①}{\le}& \sum_\ell |K'(\tilde{f}_\ell - f)|\|\mathbf{f}^\lambda - \mathbf{f}^\star\|_\infty \\ \overset{②}{\le}& \left(F_1(2.5/n, 0.2404/n) + \max_{f\in\hat{\mathcal{N}}}|K'(f)|\right)\|\mathbf{f}^\lambda - \mathbf{f}^\star\|_\infty \\ \overset{③}{\le}& (0.01241n + 0.790885n)(0.4X^\star\gamma/n) \\ \le& 0.321318X^\star\gamma,\end{aligned} \tag{A.87}$$

where ① is due to the mean value theorem with $\tilde{f}_\ell$ located between $f_\ell^\star$ and f_ℓ^λ. ② follows from Lemma A.1.5. To see this, first note that $\Delta(\{\tilde{f}_\ell\}) \geq 2.5/n$ by Lemma A.1.4. Second, $f \in \mathcal{N} = [0, 0.24/n]$ implies that

$$0 \leq |f - \tilde{f}_0| \leq |f - f_0^\star| + |f_0^\star - \tilde{f}_0| \leq 0.24/n + 0.4(10^{-3})/n = 0.2404/n.$$

We also used the definition $\hat{\mathcal{N}} = [0, 0.2404/n]$ in ②. ③ follows from the upper bound on $F_1(2.5/n, 0.2404/n)$ in Table A.1, the upper bound on $\max_{f\in\hat{\mathcal{N}}} |K'(f)|$ in Table A.3, as well as the upper bound on $\|\mathbf{f}^\lambda - \mathbf{f}^\star\|_\infty$ in Lemma 2.4.1.

Applying arguments similar to those for (A.87), we can control $\|\mathbf{D}_\ell(f, \mathbf{f}^\lambda) - \mathbf{D}_\ell(f, \mathbf{f}^\star)\|_1$ as

$$\|\mathbf{D}_\ell(f, \mathbf{f}^\lambda) - \mathbf{D}_\ell(f, \mathbf{f}^\star)\|_1 \leq (F_{\ell+1}(2.5/n, 0.2404/n) + \max_{f\in\hat{\mathcal{N}}} |K^{(\ell+1)}(f)|)\|\mathbf{f}^\lambda - \mathbf{f}^\star\|_\infty. \tag{A.88}$$

We specialize the above inequality to $\ell = 1, 2, 3$ using the upper bounds on $F_\ell(2.5/n, 0.2404/n)$ in Table A.1 and those on $\max_{f\in\hat{\mathcal{N}}} |K^{(\ell)}(f)|$ in Table A.3 to obtain

$$\begin{aligned}
\|\mathbf{D}_1(f, \mathbf{f}^\lambda) - \mathbf{D}_1(f, \mathbf{f}^\star)\|_1 \leq& (F_2(2.5/n, 0.2404/n) + \max_{f\in\hat{\mathcal{N}}} |K''(f)|)\|\mathbf{f}^\lambda - \mathbf{f}^\star\|_\infty \\
\leq& (0.05637n^2 + 3.290n^2)(0.4X^\star\gamma/n) = 1.338548nX^\star\gamma; \qquad \text{(A.89)} \\
\|\mathbf{D}_2(f, \mathbf{f}^\lambda) - \mathbf{D}_2(f, \mathbf{f}^\star)\|_1 \leq& (F_3(2.5/n, 0.2404/n) + \max_{f\in\hat{\mathcal{N}}} |K'''(f)|)\|\mathbf{f}^\lambda - \mathbf{f}^\star\|_\infty \\
\leq& (0.28838n^3 + 7.80572n^3)(0.4X^\star\gamma/n) = 3.23764n^2X^\star\gamma; \qquad \text{(A.90)} \\
\|\mathbf{D}_3(f, \mathbf{f}^\lambda) - \mathbf{D}_3(f, \mathbf{f}^\star)\|_1 \leq& (F_4(2.5/n, 0.2404/n) + \max_{f\in\hat{\mathcal{N}}} |K''''(f)|)\|\mathbf{f}^\lambda - \mathbf{f}^\star\|_\infty \\
\leq& (1.671n^4 + 29.2227n^4)(0.4X^\star\gamma/n) = 12.3575n^3X^\star\gamma. \qquad \text{(A.91)}
\end{aligned}$$

Furthermore, we can use similar arguments and Lemma A.1.5 to control $\|\mathbf{D}_\ell(f, \mathbf{f})\|_1$ for $f \in \mathcal{N}$:

$$\|\mathbf{D}_\ell(f, \mathbf{f}^\star)\|_1 \leq F_\ell(2.5/n, 0.2404/n) + \max_{f\in\hat{\mathcal{N}}} |K^{(\ell)}(f)|, \tag{A.92}$$

which specializes to

$$\begin{aligned}
\|\mathbf{D}_0(f, \mathbf{f}^\star)\|_1 \leq& F_0(2.5/n, 0.2404/n) + \max_{f\in\hat{\mathcal{N}}} |K(f)| \leq 0.00757 + 1 = 1.00757; \qquad \text{(A.93)} \\
\|\mathbf{D}_1(f, \mathbf{f}^\star)\|_1 \leq& F_1(2.5/n, 0.2404/n) + \max_{f\in\hat{\mathcal{N}}} |K'(f)| \leq 0.01241n + 0.790885n = 0.803295n; \qquad \text{(A.94)} \\
\|\mathbf{D}_2(f, \mathbf{f}^\star)\|_1 \leq& F_2(2.5/n, 0.2404/n) + \max_{f\in\hat{\mathcal{N}}} |K''(f)| \leq 0.05637n^2 + 3.290n^2 = 3.34637n^2; \qquad \text{(A.95)} \\
\|\mathbf{D}_3(f, \mathbf{f}^\star)\|_1 \leq& F_3(2.5/n, 0.2404/n) + \max_{f\in\hat{\mathcal{N}}} |K'''(f)| \leq 0.28838n^3 + 7.80572n^3 = 8.0941n^3. \qquad \text{(A.96)}
\end{aligned}$$

With these preparations, we are ready to control $\|\boldsymbol{\nu}^\star(f)^{(\ell)} - \boldsymbol{\nu}^\lambda(f)^{(\ell)}\|_1$ and $\|\boldsymbol{\nu}^\star(f)^{(\ell)}\|_1$ for $\ell = 0, 1, 2$ in Near Region. Generalizing (A.85) to the ℓth derivative of $\boldsymbol{\nu}^\star(f)$ and $\boldsymbol{\nu}^\lambda(f)$ to get

$$\begin{aligned}\|\boldsymbol{\nu}^\star(f)^{(\ell)} - \boldsymbol{\nu}^\lambda(f)^{(\ell)}\|_1 \leq & \|\mathbf{D}_{\ell+1}(f, \mathbf{f}^\lambda) - \mathbf{D}_{\ell+1}(f, \mathbf{f}^\star)\|_1 \|\operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1 \\ & + \|\mathbf{D}_{\ell+1}(f, \mathbf{f}^\lambda)\|_1 \|\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1} - \operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1 + 2\|\mathbf{D}_\ell(f, \mathbf{f}^\lambda) - \mathbf{D}_\ell(f, \mathbf{f}^\star)\|_1; \\ \|\boldsymbol{\nu}^\star(f)^{(\ell)}\|_1 \leq & \|\mathbf{D}_{\ell+1}(f, \mathbf{f}^\star)\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1}\|_1 + 2\|\mathbf{D}_\ell(f, \mathbf{f}^\star)\|_1.\end{aligned} \tag{A.97}$$

Plugging Eq. (A.87), (A.89), (A.94) and (A.86) into (A.97), we obtain

$$\begin{aligned}\|\boldsymbol{\nu}^\star(f) - \boldsymbol{\nu}^\lambda(f)\|_1 \leq & \|\mathbf{D}_1(f, \mathbf{f}^\lambda) - \mathbf{D}_1(f, \mathbf{f}^\star)\|_1 \|\operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1 \\ & + \|\mathbf{D}_1(f, \mathbf{f}^\lambda)\|_1 \|\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1} - \operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1 + 2\|\mathbf{D}_0(f, \mathbf{f}^\lambda) - \mathbf{D}_0(f, \mathbf{f}^\star)\|_1 \\ \leq & 1.338548nX^\star\gamma\frac{0.552}{n} + (0.803295n)\frac{0.552X^\star\gamma}{n} + 2(0.321318X^\star\gamma) \leq 1.82494X^\star\gamma.\end{aligned} \tag{A.98}$$

Plugging Eq. (A.89)- (A.90), (A.95) and (A.86) into (A.97), we obtain

$$\begin{aligned}\|\boldsymbol{\nu}^\star(f)' - \boldsymbol{\nu}^\lambda(f)'\|_1 \leq & \|\mathbf{D}_2(f, \mathbf{f}^\lambda) - \mathbf{D}_2(f, \mathbf{f}^\star)\|_1 \|\operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1 \\ & + \|\mathbf{D}_2(f, \mathbf{f}^\lambda)\|_1 \|\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1} - \operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1 + 2\|\mathbf{D}_1(f, \mathbf{f}^\lambda) - \mathbf{D}_1(f, \mathbf{f}^\star)\|_1 \\ \leq & 3.23764n^2X^\star\gamma\frac{0.552}{n} + (3.34637n^2)\frac{0.552X^\star\gamma}{n} + 2(1.338548nX^\star\gamma) \\ \leq & 6.31147nX^\star\gamma.\end{aligned} \tag{A.99}$$

Plugging Eq. (A.90)- (A.91), (A.96) and (A.86) into (A.97), we obtain

$$\begin{aligned}\|\boldsymbol{\nu}^\star(f)'' - \boldsymbol{\nu}^\lambda(f)''\|_1 \leq & \|\mathbf{D}_3(f, \mathbf{f}^\lambda) - \mathbf{D}_3(f, \mathbf{f}^\star)\|_1 \|\operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1 \\ & + \|\mathbf{D}_3(f, \mathbf{f}^\lambda)\|_1 \|\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1} - \operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1 + 2\|\mathbf{D}_2(f, \mathbf{f}^\lambda) - \mathbf{D}_2(f, \mathbf{f}^\star)\|_1 \\ \leq & 12.3575n^3X^\star\gamma\frac{0.552}{n} + (8.0941n^3)\frac{0.552X^\star\gamma}{n} + 2(3.23764n^2X^\star\gamma) \\ \leq & 17.7646n^2X^\star\gamma.\end{aligned} \tag{A.100}$$

Similarly, plugging Eq. (A.93)- (A.94) and (A.86) into (A.97), we have

$$\|\boldsymbol{\nu}^\star(f)\|_1 \leq \|\mathbf{D}_1(f, \mathbf{f}^\star)\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1}\|_1 + 2\|\mathbf{D}_0(f, \mathbf{f}^\star)\|_1 \leq (0.803295n)\frac{0.552}{n} + 2(1.00757) \leq 2.45856. \tag{A.101}$$

Plugging Eq. (A.94)- (A.95) and (A.86) into (A.97), we obtain

$$\|\boldsymbol{\nu}^\star(f)'\|_1 \leq \|\mathbf{D}_2(f, \mathbf{f}^\star)\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1}\|_1 + 2\|\mathbf{D}_1(f, \mathbf{f}^\star)\|_1 \leq (3.34637n^2)\frac{0.552}{n} + 2(0.803295n) \leq 3.4538n. \tag{A.102}$$

Finally, plugging Eq. (A.95)- (A.96) and (A.86) into (A.97), we arrive at

$$\|\boldsymbol{\nu}^\star(f)''\|_1 \le \|\mathbf{D}_3(f,\mathbf{f}^\star)\,\mathrm{diag}(\mathbf{c}^\star)\mathbf{S}^{-1}\|_1 + 2\|\mathbf{D}_2(f,\mathbf{f}^\star)\|_1 \le (8.0941n^3)\frac{0.552}{n} + 2(3.34637n^2) \le 11.1607n^2. \tag{A.103}$$

We are now ready to control the pointwise distance between $Q^{\star(\ell)}(f)$ and $Q^{\lambda(\ell)}(f)$ using

$$|Q^{\star(\ell)}(f) - Q^{\lambda(\ell)}(f)| \le |\boldsymbol{\nu}^\lambda(f)^{(\ell)}\Xi^\lambda\boldsymbol{\rho}^\lambda - \boldsymbol{\nu}^\star(f)^{(\ell)}\Xi^\star\boldsymbol{\rho}^\star|, \quad \ell = 0,1,2. \tag{A.104}$$

Plugging Eq. (A.98)- (A.99), (A.75)- (A.83) and (A.84) to (A.104) with $\ell = 0$, we obtain for $f \in \mathcal{N}$

$$\begin{aligned}
&|Q^\star(f) - Q^\lambda(f)| \\
\le&\|\boldsymbol{\nu}^\star(f)\|_1\|\Xi^\star\|_{\infty,\infty}\|\boldsymbol{\rho}^\star - \boldsymbol{\rho}^\lambda\|_1 + \|\boldsymbol{\nu}^\star(f)\|_1\|\Xi^\star - \Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty + \|\boldsymbol{\nu}^\star(f) - \boldsymbol{\nu}^\lambda(f)\|_1\|\Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty \\
\le&(2.45856)(1.03172)(2X^\star\gamma) + (2.45856)(8.81222X^\star B^\star\gamma) + (1.82494X^\star\gamma)(1.09363) \le 28.7343X^\star B^\star\gamma.
\end{aligned}$$

Plugging Eq. (A.100)- (A.101), (A.75)- (A.83) and (A.84) to (A.104) with $\ell = 1$, we obtain for $f \in \mathcal{N}$

$$\begin{aligned}
&|Q^\star(f)' - Q^\lambda(f)'| \\
\le&\|\boldsymbol{\nu}^\star(f)'\|_1\|\Xi^\star\|_{\infty,\infty}\|\boldsymbol{\rho}^\star - \boldsymbol{\rho}^\lambda\|_1 + \|\boldsymbol{\nu}^\star(f)'\|_1\|\Xi^\star - \Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty + \|\boldsymbol{\nu}^\star(f)' - \boldsymbol{\nu}^\lambda(f)'\|_1\|\Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty \\
\le&(3.4538n)(1.03172)(2X^\star\gamma) + (3.4538n)(8.81222X^\star B^\star\gamma) + (6.31147nX^\star\gamma)(1.09363) \le 44.4648nX^\star B^\star\gamma.
\end{aligned}$$

Finally, plugging Eq. (A.102)- (A.103), (A.75)- (A.83) and (A.84) to (A.104) with $\ell = 2$, we get for $f \in \mathcal{N}$

$$\begin{aligned}
&|Q^\star(f)'' - Q^\lambda(f)''| \\
\le&\|\boldsymbol{\nu}^\star(f)''\|_1\|\Xi^\star\|_{\infty,\infty}\|\boldsymbol{\rho}^\star - \boldsymbol{\rho}^\lambda\|_1 + \|\boldsymbol{\nu}^\star(f)''\|_1\|\Xi^\star - \Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty + \|\boldsymbol{\nu}^\star(f)'' - \boldsymbol{\nu}^\lambda(f)''\|_1\|\Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty \\
\le&(11.1607n^2)(1.03172)(2X^\star\gamma) + (11.1607n^2)(8.81222X^\star B^\star\gamma) + (17.7646n^2X^\star\gamma)(1.09363) \le 140.808n^2X^\star B^\star\gamma.
\end{aligned}$$

A.7.2 Middle Region Analysis

We continue with bounding the pointwise distance between $Q^\star(f)$ and $Q^\lambda(f)$ in Middle Region $\mathcal{M} = [0.24/n, 0.75/n]$. We start with controlling $\|\mathbf{D}_\ell(f,\mathbf{f}^\lambda) - \mathbf{D}_\ell(f,\mathbf{f}^\star)\|_1$ and $|\mathbf{D}_\ell(f,\mathbf{f}^\star)\|_1$ for $\ell = 0,1$. First note when $f \in \mathcal{M} = [0.24/n, 0.75/n]$, we have

$$\begin{aligned}
(a)\ |f - \tilde{f}_0| \le&|f - f_0^\star| + |f_0^\star - \tilde{f}_0| \le 0.75/n + 0.0004/n = 0.7504/n, \\
(b)\ |f - \tilde{f}_0| \ge&|f - f_0^\star| - |f_0^\star - \tilde{f}_0| \ge 0.24/n - 0.0004/n = 0.2396/n.
\end{aligned}$$

Denote $\hat{\mathcal{M}} = [0.2396/n, 0.7504/n]$. We combine the upper bounds on $F_\ell(2.5/n, 0.7504/n)$ in Table A.1 and the upper bounds on $\max_{f\in\hat{\mathcal{M}}}|K^{(\ell)}(f)|$ in Table A.3 to get

$$\begin{aligned}
\|\mathbf{D}_0(f,\mathbf{f}^\lambda) - \mathbf{D}_0(f,\mathbf{f}^\star)\|_1 \le&(F_1(2.5/n, 0.7504/n) + \max_{f\in\hat{\mathcal{M}}}|K'(f)|)\|\mathbf{f}^\lambda - \mathbf{f}^\star\|_\infty \\
\le&(0.01454n + 2.46872n)(0.4X^\star\gamma/n) = 0.993304X^\star\gamma;
\end{aligned} \tag{A.105}$$

$$\begin{aligned}\|\mathbf{D}_1(f,\mathbf{f}^\lambda)-\mathbf{D}_1(f,\mathbf{f}^\star)\|_1 \leq & (F_2(2.5/n,0.7504/n)+\max_{f\in\hat{\mathcal{M}}}|K''(f)|)\|\mathbf{f}^\lambda-\mathbf{f}^\star\|_\infty \\ \leq & (0.12675n^2+3.290n^2)(0.4X^\star\gamma/n)=1.36670nX^\star\gamma.\end{aligned} \tag{A.106}$$

In a similar manner, we use Lemma A.1.5 to control $\|\mathbf{D}_\ell(f,\mathbf{f})\|_1$ as follows

$$\|\mathbf{D}_0(f,\mathbf{f})\|_1 \leq F_0(2.5/n,0.7504/n)+\max_{f\in\hat{\mathcal{M}}}|K(f))| \leq 0.00772+0.90951=0.91723; \tag{A.107}$$

$$\|\mathbf{D}_1(f,\mathbf{f})\|_1 \leq F_1(2.5/n,0.7504/n)+\max_{f\in\hat{\mathcal{M}}}|K'(f))| \leq 0.01454n+2.46872n=2.48326n. \tag{A.108}$$

To control $\|\boldsymbol{\nu}^\star(f)-\boldsymbol{\nu}^\lambda(f)\|_1$ and $\|\boldsymbol{\nu}^\star(f)\|_1$ in the Middle Region, we plug Eq. (A.105)- (A.108) into (A.97) to get

$$\begin{aligned}\|\boldsymbol{\nu}^\star(f)-\boldsymbol{\nu}^\lambda(f)\|_1 \leq & \|\mathbf{D}_1(f,\mathbf{f}^\lambda)-\mathbf{D}_1(f,\mathbf{f}^\star)\|_1\|\operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1+\|\mathbf{D}_1(f,\mathbf{f}^\lambda)\|_1\|\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1}-\operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1 \\ & +2\|\mathbf{D}_0(f,\mathbf{f}^\lambda)-\mathbf{D}_0(f,\mathbf{f}^\star)\|_1 \\ \leq & 1.36670nX^\star\gamma\frac{0.552}{n}+(2.48326n)\frac{0.552X^\star\gamma}{n}+2(0.993304X^\star\gamma)\leq 4.11179X^\star\gamma; \end{aligned} \tag{A.109}$$

$$\|\boldsymbol{\nu}^\star(f)\|_1 \leq \|\mathbf{D}_1(f,\mathbf{f}^\star)\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1}\|_1+2\|\mathbf{D}_0(f,\mathbf{f}^\star)\|_1 \leq (2.48326n)\frac{0.552}{n}+2(0.91723)\leq 3.20522. \tag{A.110}$$

Finally, we control $|Q^\star(f)-Q^\lambda(f)|$ in Middle Region by plugging Eq. (A.109)- (A.110), (A.75)- (A.83) and (A.84) to (A.104) with $\ell=0$ to get

$$\begin{aligned}|Q^\star(f)-Q^\lambda(f)| \leq & \|\boldsymbol{\nu}^\star(f)\|_1\|\Xi^\star\|_{\infty,\infty}\|\boldsymbol{\rho}^\star-\boldsymbol{\rho}^\lambda\|_1+\|\boldsymbol{\nu}^\star(f)\|_1\|\Xi^\star-\Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty \\ & +\|\boldsymbol{\nu}^\star(f)-\boldsymbol{\nu}^\lambda(f)\|_1\|\Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty \\ \leq & (3.20522)(1.03172)(2X^\star\gamma)+(3.20522)(8.81222X^\star B^\star\gamma)+(4.11179X^\star\gamma)1.09363 \\ \leq & 39.3557X^\star B^\star\gamma,\ f\in\mathcal{M}.\end{aligned}$$

A.7.3 Far Region Analysis

Lastly, we bound the pointwise distance between $Q^\star(f)$ and $Q^\lambda(f)$ in Far Region $\mathcal{F}=[0.75/n, f_1^\star/2]$. Again, we start with controlling $\|\mathbf{D}_\ell(f,\mathbf{f}^\lambda)-\mathbf{D}_\ell(f,\mathbf{f}^\star)\|_1$ and $|\mathbf{D}_\ell(f,\mathbf{f}^\star)\|_1$ for $\ell=0,1$. First note when $f\in\mathcal{F}=[0.75/n, f_1^\star/2]$, we have

$$\begin{aligned}&(a)\ f-\tilde{f}_0 \geq f-f_0^\star-|f_0^\star-\tilde{f}_0| \geq 0.75/n-0.0004/n=0.74996/n, \\ &(b)\ \tilde{f}_1-f \geq -|\tilde{f}_1-f_1^\star|+f_1^\star-f \geq -0.0004n+f_1^\star/2 \geq -0.0004/n+(2.5009/n)/2 \geq 1.25/n.\end{aligned}$$

Further note that $\{\tilde{f}_\ell\}$ satisfies the separation condition that $\Delta(\{\tilde{f}_\ell\})\geq 2.5/n$. Then, following from Lemma A.1.3 and the upper bounds on $W_\ell(0.74996/n, 1.25/n)$ in Table A.2, we have

$$\begin{aligned}\|\mathbf{D}_0(f,\mathbf{f}^\lambda)-\mathbf{D}_0(f,\mathbf{f}^\star)\|_1 \leq& \|\mathbf{D}_1(\tilde{\mathbf{f}},f)\|_1\|\mathbf{f}^\lambda-\mathbf{f}^\star\|_\infty \\ \leq& W_1(0.74996/n,1.25/n)\|\mathbf{f}^\lambda-\mathbf{f}^\star\|_\infty \leq 5.2265n(0.4X^\star\gamma/n)=2.0906X^\star\gamma; \quad \text{(A.111)}\\ \|\mathbf{D}_1(f,\mathbf{f}^\lambda)-\mathbf{D}_1(f,\mathbf{f}^\star)\|_1 \leq& W_2(0.74996/n,1.25/n)\|\mathbf{f}^\lambda-\mathbf{f}^\star\|_\infty \leq 48.033n^2(0.4X^\star\gamma/n)=19.2132nX^\star\gamma. \quad \text{(A.112)}\end{aligned}$$

Similarly, we can use Lemma A.1.3 to control $\|\mathbf{D}_\ell(f,\mathbf{f})\|_1$ for $\ell=0,1$ and $f\in\mathcal{F}$:

$$\begin{aligned}\|\mathbf{D}_0(f,\mathbf{f}^\star)\|_1 \leq& W_0(0.74996/n,1.25/n)\leq 0.71059;\\ \|\mathbf{D}_1(f,\mathbf{f}^\star)\|_1 \leq& W_1(0.74996/n,1.25/n)\leq 5.2265n.\end{aligned} \quad \text{(A.113)}$$

Directly plugging Eq. (A.111)- (A.113) into (A.97), we arrive at

$$\begin{aligned}\|\boldsymbol{\nu}^\star(f)-\boldsymbol{\nu}^\lambda(f)\|_1 \leq& \|\mathbf{D}_1(f,\mathbf{f}^\lambda)-\mathbf{D}_1(f,\mathbf{f}^\star)\|_1\|\operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1+\|\mathbf{D}_1(f,\mathbf{f}^\lambda)\|_1\|\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1}-\operatorname{diag}(\mathbf{c}^\lambda)\mathbf{S}^{-1}\|_1\\ &+2\|\mathbf{D}_0(f,\mathbf{f}^\lambda)-\mathbf{D}_0(f,\mathbf{f}^\star)\|_1\\ \leq& 19.2132nX^\star\gamma\frac{0.552}{n}+(5.2265n)\frac{0.552X^\star\gamma}{n}+2(2.0906X^\star\gamma)\leq 17.6720X^\star\gamma; \quad \text{(A.114)}\\ \|\boldsymbol{\nu}^\star(f)\|_1 \leq& \|\mathbf{D}_1(f,\mathbf{f}^\star)\operatorname{diag}(\mathbf{c}^\star)\mathbf{S}^{-1}\|_1+2\|\mathbf{D}_0(f,\mathbf{f}^\star)\|_1\leq(5.2265n)\frac{0.552}{n}+2(0.71059)\leq 4.30621. \quad \text{(A.115)}\end{aligned}$$

As a final step, we control $|Q^\star(f)-Q^\lambda(f)|$ in Far Region by plugging Eq. (A.114)- (A.115) and (A.75)- (A.84) to (A.104) to get

$$\begin{aligned}|Q^\star(f)-Q^\lambda(f)| \leq& \|\boldsymbol{\nu}^\star(f)\|_1\|\Xi^\star\|_{\infty,\infty}\|\boldsymbol{\rho}^\star-\boldsymbol{\rho}^\lambda\|_1+\|\boldsymbol{\nu}^\star(f)\|_1\|\Xi^\star-\Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty\\ &+\|\boldsymbol{\nu}^\star(f)-\boldsymbol{\nu}^\lambda(f)\|_1\|\Xi^\lambda\|_{\infty,\infty}\|\boldsymbol{\rho}^\lambda\|_\infty\\ \leq& (4.30621)(1.03172)(2X^\star\gamma)+(4.30621)(8.81222X^\star B^\star\gamma)+(17.6720X^\star\gamma)(1.09363)\\ \leq& 66.1596X^\star B^\star\gamma,\ f\in\mathcal{F}.\end{aligned}$$

This concludes the proof of Lemma 2.4.4. □

A.8 Proof of Lemma 2.4.5

Lemma A.8.1 (Lemma 2.4.5)**.** *Under the settings of Lemma 2.4.2, let $\hat{Q}$ and Q^λ be the dual polynomials corresponding to $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\lambda$, respectively. Then the pointwise distances between $Q^\lambda(f)$ and $\hat{Q}(f)$ and their derivatives are bounded:*

$$\begin{aligned}&|\hat{Q}(f)-Q^\lambda(f)|\leq 82.5975B^\star/X^\star,\ f\in\mathcal{N}, && |\hat{Q}(f)-Q^\lambda(f)|\leq 114.323B^\star/X^\star,\ f\in\mathcal{M},\\ &|\hat{Q}(f)'-Q^{\lambda'}(f)|\leq 180.283nB^\star/X^\star,\ f\in\mathcal{N}, && |\hat{Q}(f)-Q^\lambda(f)|\leq 162.903B^\star/X^\star,\ f\in\mathcal{F},\\ &|\hat{Q}(f)''-Q^{\lambda''}(f)|\leq 758.404n^2B^\star/X^\star,\ f\in\mathcal{N}.\end{aligned}$$

Proof. The expressions $\mathbf{q}^\lambda=\frac{\mathbf{x}^\star-\mathbf{x}^\lambda}{\lambda}$ and $\hat{\mathbf{q}}=\frac{\mathbf{y}-\hat{\mathbf{x}}}{\lambda}$ lead to

$$\hat{\mathbf{q}} - \mathbf{q}^\lambda = \frac{(\mathbf{y} - \hat{\mathbf{x}}) - (\mathbf{x}^\star - \mathbf{x}^\lambda)}{\lambda} = \frac{\mathbf{w}}{\lambda} + \frac{\mathbf{x}^\lambda - \hat{\mathbf{x}}}{\lambda},$$

implying

$$|Q^\lambda(f) - \hat{Q}(f)| \leq \underbrace{\frac{|\mathbf{a}(f)^H \mathbf{Z}\mathbf{w}|}{\lambda}}_{\Pi_1(f)} + \underbrace{\frac{|\mathbf{a}(f)^H \mathbf{Z}(\mathbf{x}^\lambda - \hat{\mathbf{x}})|}{\lambda}}_{\Pi_2(f)}. \tag{A.116}$$

This separates the distance between $Q^\lambda(f)$ and $\hat{Q}(f)$ into two parts: one is $\Pi_1(f)$ associated with the dual atomic norm of the Gaussian noise $\mathbf{w}$ whose upper bounds were developed in Appendix A.2; the other is $\Pi_2(f)$ corresponding to the dual atomic norm of $\mathbf{x}^\lambda - \hat{\mathbf{x}}$. The latter quantity can be bounded by similar arguments as controlling $|\mathbf{a}(f)^H \mathbf{Z}(\mathbf{x}^\lambda - \hat{\mathbf{x}})|$ in Lemma 2.4.4.

Bounding $\Pi_1(f)$.

Combining Eq. (A.32)- (A.34), we can upperbound $\Pi_1(f), \Pi(f)'$ and $\Pi_1(f)''$ with high probability (at least $1 - 1/n^2$) for all $f \in \mathbb{T}$:

$$\begin{aligned} &\Pi_1(f) \leq 6.534\gamma_0/\lambda \leq 10.115/X^\star; \\ &\Pi_1(f)' \leq 41.052n\gamma_0/\lambda \leq 63.458n/X^\star; \\ &\Pi_1(f)'' \leq 257.94n^2\gamma_0/\lambda \leq 399.288n^2/X^\star, \end{aligned} \tag{A.117}$$

where we used $\lambda = 0.646X^\star\gamma_0$.

Bounding $\Pi_2(f)$.

$$\begin{aligned} \Pi_2(f) &= \frac{1}{\lambda}|\mathbf{D}_0(f, \mathbf{f}^\lambda)\mathbf{c}^\lambda - \mathbf{D}_0(f, \hat{\mathbf{f}})\hat{\mathbf{c}}| \\ &\leq \frac{1}{\lambda}\|\mathbf{D}_0(f, \mathbf{f}^\lambda) - \mathbf{D}_0(f, \hat{\mathbf{f}})\|_1\|\mathbf{c}^\lambda\|_\infty + \frac{1}{\lambda}\|\mathbf{D}_0(f, \hat{\mathbf{f}})\|_1\|\hat{\mathbf{c}} - \mathbf{c}^\lambda\|_\infty \\ &\overset{①}{\leq} \frac{1}{\lambda}\|\mathbf{D}_1(f, \tilde{\mathbf{f}})\|_1\|\mathbf{f}^\lambda - \hat{\mathbf{f}}\|_\infty\|\mathbf{c}^\lambda\|_\infty + \frac{1}{\lambda}\|\mathbf{D}_0(f, \hat{\mathbf{f}})\|_1\|\hat{\mathbf{c}} - \mathbf{c}^\lambda\|_\infty \\ &\overset{②}{\leq} \frac{c^\lambda_{\max}}{0.646X^\star\gamma_0}\left(\frac{0.4(35.2)\gamma}{n}\|\mathbf{D}_1(f, \tilde{\mathbf{f}})\|_1 + 35.2\gamma\|\mathbf{D}_0(f, \hat{\mathbf{f}})\|_1\right) \\ &\overset{③}{\leq} \frac{B^\star(1 + X^\star\gamma)}{0.646X^\star}\left(\frac{14.08}{n}\|\mathbf{D}_1(f, \tilde{\mathbf{f}})\|_1 + 35.2\|\mathbf{D}_0(f, \hat{\mathbf{f}})\|_1\right), \end{aligned}$$

where ① follows from the mean value theorem. For ② to hold, first note that $\lambda = 0.646X^\star\gamma_0$ and $\hat{\boldsymbol{\theta}} \in N^\lambda$ by Lemma 2.4.2. Then, we can upperbound $\|\hat{\mathbf{c}} - \mathbf{c}^\lambda\|_\infty$ as

$$\|\hat{\mathbf{c}} - \mathbf{c}^\lambda\|_\infty = \frac{|\hat{c}_j - c^\lambda_j|}{|c^\lambda_j|}|c^\lambda_j| \leq (35.2\gamma)c^\lambda_{\max},$$

where the equality follows by assuming the ℓ_∞ norm is achieved by the jth row and the inequality follows by changing $X^\star$ to 35.2 in (A.10) and defining $c^\lambda_{\max} := \max_j |c^\lambda_j|$. ③ follows from $\gamma_0 = \gamma c^\star_{\min}$ and

$$\frac{c^{\lambda}_{\max}}{c^{\star}_{\min}} = B^{\star}\frac{|c^{\lambda}_j|}{c^{\star}_{\max}} \leq B^{\star}\frac{|c^{\lambda}_j|}{|c^{\star}_j|} \leq B^{\star}(1+X^{\star}\gamma).$$

As a consequence, to control $\Pi_2(f)$, it reduces to bounding $\|\mathbf{D}_{\ell}(f,\tilde{\mathbf{f}})\|_1$ and $\|\mathbf{D}_{\ell}(f,\hat{\mathbf{f}})\|_1$. For this purpose, we first note that $\{\Delta(\tilde{T}),\Delta(\hat{T})\} \geq 2.5/n$ by Lemma A.1.4. Second, by

$$\|\tilde{\mathbf{f}}-\mathbf{f}^{\star}\|_{\infty} \overset{①}{\leq} \|\hat{\mathbf{f}}-\mathbf{f}^{\star}\|_{\infty} \overset{②}{\leq} 0.4(X^{\star}+35.2)\gamma \overset{③}{\leq} 0.0004/n + 1.408\times 10^{-6}/n = 0.000401408/n,$$

where ① follows from the length of subinterval is no larger than the whole one. ② follows from Eq. (A.10) and ③ follows from the SNR condition (2.10). Thus, we can follow the same arguments that lead to Eq. (A.93)- (A.94) for Near Region, Eq. (A.107)- (A.108) for Middle Region, and Eq. (A.113) for Far Region to develop bounds on $\|\mathbf{D}_{\ell}(f,\hat{\mathbf{f}})\|_1$.

To have a concrete idea, we first show how to control $\|\mathbf{D}_{\ell}(f,\hat{\mathbf{f}})\|_1$ since the upper bounds for $\|\mathbf{D}_{\ell}(f,\tilde{\mathbf{f}})\|_1$ then follows by $\|\tilde{\mathbf{f}}-\mathbf{f}^{\star}\|_{\infty} \leq \|\hat{\mathbf{f}}-\mathbf{f}^{\star}\|_{\infty}$. First, consider $f \in \mathcal{N}$. Then we have

$$0 \leq |\hat{f}_0 - f| \leq |\hat{f}_0 - f^{\star}_0| + |f^{\star}_0 - f| \leq 0.000401408/n + 0.24/n \leq 0.240401408/n.$$

With some abuse of notation, we denote $\hat{\mathcal{N}} := [0, 0.240401408/n]$. Second, consider $f \in \mathcal{M}$. Then we have

$$\begin{aligned}(a)\ |f-\hat{f}_0| \leq& |f-f^{\star}_0| + |f^{\star}_0 - \hat{f}_0| \leq 0.75/n + 0.000401408/n = 0.750401408/n;\\ (b)\ |f-\hat{f}_0| \geq& |f-f^{\star}_0| - |f^{\star}_0 - \hat{f}_0| \geq 0.24/n - 0.000401408/n. = 0.239598592/n.\end{aligned}$$

Denote $\hat{\mathcal{M}} = [0.2396/n, 0.7504/n]$. At last, we consider $f \in \mathcal{F} = [0.75/n, f^{\star}_1/2]$:

$$\begin{aligned}&(a)\ f-\hat{f}_0 \geq f - f^{\star}_0 - |f^{\star}_0 - \hat{f}_0| \geq 0.75/n - 0.000401408/n = 0.749598592/n;\\ &(b)\ \hat{f}_1 - f \geq -|\hat{f}_1 - f^{\star}_1| + f^{\star}_1 - f \geq -0.000401408/n + f^{\star}_1/2 \geq -0.000401408/n + (2.5009/n)/2 \geq 1.25/n.\end{aligned}$$

Hence we can define $\hat{\mathcal{F}} := [0.749598592/n, 1.25/n]$. Furthermore, we remark that those numerical upper bounds in Table A.1- Table A.3 do not change when evaluated for the newly defined intervals $\hat{\mathcal{N}}$, $\hat{\mathcal{M}}$ and $\hat{\mathcal{F}}$.

Finally, by directly plugging the upper bounds of $\|\mathbf{D}_{\ell}(f,\mathbf{f})\|_{\infty}$ in (A.93)- (A.94) for Near Region, (A.107)- (A.108) for Middle Region, and equation (A.113) for Far Region, it follows that

$$\Pi_2(f) \leq \frac{B^{\star}}{X^{\star}}\begin{cases}\frac{1.001}{0.646}(\frac{14.08}{n}(0.803295n)+35.2(1.00757)) \leq 72.4825\frac{B^{\star}}{X^{\star}},\ f\in\mathcal{N};\\ \frac{1.001}{0.646}(\frac{14.08}{n}(2.48326n)+35.2(0.91723)) \leq 104.208\frac{B^{\star}}{X^{\star}},\ f\in\mathcal{M};\\ \frac{1.001}{0.646}(\frac{14.08}{n}(5.2265n)+35.2(0.71059)) \leq 152.788\frac{B^{\star}}{X^{\star}},\ f\in\mathcal{F}.\end{cases} \tag{A.118}$$

Similarly, from (A.94)- (A.96), we have an upper bound on $\Pi_2(f)'$ and $\Pi_2(f)''$ as follows

$$\Pi_2(f)' \leq \frac{B^\star(1+X^\star\gamma)}{0.646X^\star}\left(\frac{14.08}{n}\|\mathbf{D}_2(f,\tilde{\mathbf{f}})\|_1 + 35.2\|\mathbf{D}_1(f,\hat{\mathbf{f}})\|_1\right)$$
$$\leq \frac{B^\star}{X^\star}\frac{1.001}{0.646}\left(\frac{14.08}{n}(3.34637n^2) + 35.2(0.803295n)\right) \leq 116.825n\frac{B^\star}{X^\star},\ f \in \mathcal{N}; \tag{A.119}$$

$$\Pi_2(f)'' \leq \frac{B^\star(1+X^\star\gamma)}{0.646X^\star}\left(\frac{14.08}{n}\|\mathbf{D}_3(f,\tilde{\mathbf{f}})\|_1 + 35.2\|\mathbf{D}_2(f,\hat{\mathbf{f}})\|_1\right)$$
$$\leq \frac{B^\star}{X^\star}\frac{1.001}{0.646}\left(\frac{14.08}{n}(8.0941n^3) + 35.2(3.34637n^2)\right) \leq 359.116n^2\frac{B^\star}{X^\star},\ f \in \mathcal{N}. \tag{A.120}$$

Combining (A.117)- (A.120) for $\Pi_1(f)$ and $\Pi_2(f)$, we can control $|\hat{Q}^{(\ell)}(f) - Q^{\lambda^{(\ell)}}(f)|$ in Near region $f \in \mathcal{N}$ as follows

$$|\hat{Q}(f) - Q^\lambda(f)| \leq (10.115 + 72.4825)B^\star/X^\star = 82.5975B^\star/X^\star,\ f \in \mathcal{N};$$
$$|\hat{Q}(f)' - Q^{\lambda'}(f)| \leq (63.458n + 116.825n)B^\star/X^\star = 180.283nB^\star/X^\star,\ f \in \mathcal{N};$$
$$|\hat{Q}(f)'' - Q^{\lambda''}(f)| \leq (399.288n^2 + 359.116n^2)B^\star/X^\star = 758.404n^2B^\star/X^\star,\ f \in \mathcal{N}.$$

For the case of Middle Region and Far Region, we can upperbound them as:

$$|\hat{Q}(f) - Q^\lambda(f)| \leq (10.115 + 104.208)B^\star/X^\star = 114.323B^\star/X^\star,\ f \in \mathcal{M};$$
$$|\hat{Q}(f) - Q^\lambda(f)| \leq (10.115 + 152.788)B^\star/X^\star = 162.903B^\star/X^\star,\ f \in \mathcal{F}.$$

This completes the proof of Lemma 2.4.5. □

A.9 Proof of Proposition 2.4.1

Proposition A.9.1 (Proposition 2.4.1)**.** *Let the decomposition* $\hat{\mathbf{x}} = \sum_{\ell=1}^{\hat{k}} \hat{c}_\ell \mathbf{a}(\hat{f}_\ell)$ *with distinct frequencies* $\hat{T} = \{\hat{f}_\ell\} \subset \mathbb{T}$ *and nonzero coefficients* $\{\hat{c}_\ell\}$ *and set* $\hat{\mathbf{q}} = (\mathbf{y} - \hat{\mathbf{x}})/\lambda$*. Suppose the corresponding dual polynomial* $\hat{Q}(f) = \mathbf{a}(f)^H\mathbf{Z}\hat{\mathbf{q}}$ *satisfies the following* Bounded Interpolation Property (BIP)*:*

$$\hat{Q}(\hat{f}_\ell) = \operatorname{sign}(\hat{c}_\ell), \ell = 1, \ldots, \hat{k} \ \text{(Interpolation)};$$
$$|\hat{Q}(f)| < 1, \forall f \notin \hat{T} \ \text{(Boundedness)};$$

then $\hat{\mathbf{x}}$ *and* $\hat{\mathbf{q}}$ *are the unique primal-dual optimal solutions to* (2.8) *and* (2.16)*, that is,* $\hat{\mathbf{x}} = \mathbf{x}^{\text{glob}}$ *and* $\hat{\mathbf{q}} = \mathbf{q}^{\text{glob}}$*. Here the operation* $\operatorname{sign}(c) := c/|c|$ *for a nonzero complex number and applies entry-wise to a vector.*

Proof. The uniqueness follows from the strongly convex quadratic term in (2.8). We next show the primal optimality of $\hat{\mathbf{x}}$ and the dual optimality of $\hat{\mathbf{q}}$ by establishing strong duality. First, $\hat{\mathbf{q}}$ is feasible to the dual program (2.16) because of the BIP property. Second, we have the following chain of inequalities:

$$\begin{aligned}\text{value of } (2.16) &= \frac{1}{2}\|\mathbf{y}\|_{\mathbf{Z}}^2 - \frac{1}{2}\|\mathbf{y} - \lambda\hat{\mathbf{q}}\|_{\mathbf{Z}}^2 \\ &= \frac{1}{2}\|\lambda\hat{\mathbf{q}}\|_{\mathbf{Z}}^2 + \lambda\mathbb{R}\{\hat{\mathbf{x}}^H\mathbf{Z}\hat{\mathbf{q}}\} \\ &= \frac{1}{2}\|\mathbf{y} - \hat{\mathbf{x}}\|_{\mathbf{Z}}^2 + \lambda\|\hat{\mathbf{c}}\|_1 \\ &\geq \frac{1}{2}\|\mathbf{y} - \hat{\mathbf{x}}\|_{\mathbf{Z}}^2 + \lambda\|\hat{\mathbf{x}}\|_{\mathcal{A}} = \text{value of } (2.8),\end{aligned}$$

where the second line follows by plugging $\mathbf{y} = \hat{\mathbf{x}} + \lambda\hat{\mathbf{q}}$; the third line holds due to the Interpolation property; and the last line holds since $\|\hat{\mathbf{x}}\|_{\mathcal{A}} \leq \|\hat{\mathbf{c}}\|_1$ by (2.7). Since the weak duality theorem ensures that the other direction of the inequality always holds, we obtain strong duality. As a consequence, $\hat{\mathbf{x}}$ and $\hat{\mathbf{q}}$ achieve primal optimality and dual optimality, respectively. This means $\hat{\mathbf{x}} = \mathbf{x}^{\text{glob}}, \hat{\mathbf{q}} = \mathbf{q}^{\text{glob}}$ due to uniqueness of the solutions. □

A.10 Proof of Corollary 2.2.1

Corollary A.10.1 (Corollary 2.2.1)**.** *Under the same setup as in Theorem 2.2.1, with probability at least* $1 - \frac{1}{n^2}$*, the frequencies and coefficients estimated by the atomic norm regularized minimization* (2.8) *constitute a global optimum of the* ℓ_1*-regularized nonlinear least-squares program* (2.15)*.*

Proof. Denote by $F(\mathbf{x})$ the objective functions for (2.8) and $G(\mathbf{f}, \mathbf{c})$ for (2.15). Assume $(\mathbf{f}^{\text{non}}, \mathbf{c}^{\text{non}})$ is a global optimum for (2.15) with $\mathbf{x}^{\text{non}} = \mathbf{A}(\mathbf{f}^{\text{non}})\mathbf{c}^{\text{non}}$, and $\mathbf{x}^{\text{glob}} = \mathbf{A}(\mathbf{f}^{\text{glob}})\mathbf{c}^{\text{glob}}$ is the global optimum of (2.8). Then

$$F(\mathbf{x}^{\text{glob}}) \leq F(\mathbf{x}^{\text{non}}) \leq G(\mathbf{f}^{\text{non}}, \mathbf{c}^{\text{non}}) \leq G(\mathbf{f}^{\text{glob}}, \mathbf{c}^{\text{glob}}), \tag{A.121}$$

where the first inequality uses the optimality of $\mathbf{x}^{\text{glob}}$ to (2.8); the second inequality follows from $\|\mathbf{x}^{\text{non}}\|_{\mathcal{A}} \leq \|\mathbf{c}_\ell^{\text{non}}\|_1$ by (2.7); and the last inequality follows from the optimality of $(\mathbf{f}^{\text{non}}, \mathbf{c}^{\text{non}})$ to (2.15). On the other hand, recognize that $\|\mathbf{x}^{\text{glob}}\|_{\mathcal{A}} = \|\mathbf{c}_\ell^{\text{glob}}\|_1$ since $\{f_\ell^{\text{glob}}\}$ satisfies the separation condition (revealed by Lemma A.1.4 in Appendix A.1). This leads to $G(\mathbf{f}^{\text{glob}}, \mathbf{c}^{\text{glob}}) = F(\mathbf{x}^{\text{glob}})$. Therefore, all inequalities in (A.121) become equalities and hence $G(\mathbf{f}^{\text{non}}, \mathbf{c}^{\text{non}}) = G(\mathbf{f}^{\text{glob}}, \mathbf{c}^{\text{glob}})$. This implies the global optimality of $(\mathbf{f}^{\text{glob}}, \mathbf{c}^{\text{glob}})$ for the nonconvex program (2.15).

□

A.11 Proof of Lemma A.1.4

Lemma A.11.1 (Lemma A.1.4)**.** *Let the separation condition* (2.9) *and the SNR condition* (2.10) *hold. Then both the frequencies* $T^\lambda = \{f_\ell^\lambda\}$ *returned by the first fixed point map* (2.19) *and the frequencies* $\hat{T} = \{\hat{f}_\ell\}$ *generated by the second fixed point map* (2.21) *have minimal separations at least* $2.5/n$*. Furthermore, the intermediate frequencies defined by* $\tilde{T} = \{\tilde{f}_\ell\}_{\ell=1}^k$ *with each* $\tilde{f}_\ell \in [f_\ell^\star, f_\ell^\lambda]$ *or* $[f_\ell^\lambda, f_\ell^\star]$ *and the second intermediate frequencies* $\tilde{T}^\lambda := \{\tilde{f}_\ell\}_{\ell=1}^k$ *with each* $\tilde{f}_\ell \in [f_\ell^\lambda, \hat{f}_\ell]$ *or* $[\hat{f}_\ell, f_\ell^\lambda]$ *also have minimal separations at least* $2.5/n$*:*

$$\min\{\Delta(T^\lambda), \Delta(\tilde{T}), \Delta(\hat{T}), \Delta(\tilde{T}^\lambda)\} \geq 2.5/n.$$

Proof. First of all, from Lemma 2.4.1, we have $\boldsymbol{\theta}^\lambda \in \mathcal{N}^\star$, implying $\|\mathbf{f}^\lambda - \mathbf{f}^\star\|_\infty \leq 0.4X^\star B^\star \gamma/n$ by Eq. (A.10) and by Lemma 2.4.2, we obtain that $\hat{\boldsymbol{\theta}} \in \mathcal{N}^\lambda$, which implies $\|\hat{\mathbf{f}} - \mathbf{f}^\lambda\|_\infty \leq 0.4(35.2)B^\star\gamma/n$ by Eq. (A.10). More precisely, we bound $\Delta(T^\lambda)$ as

$$\begin{aligned}
\Delta(T^\lambda) &\overset{①}{=} \min_{i\neq j} |f_i^\lambda - f_j^\lambda| \\
&= \min_{i\neq j} |f_i^\lambda - f_i^\star + f_i^\star - f_j^\star + f_j^\star - f_j^\lambda| \\
&\overset{②}{\geq} \min_{i\neq j} |f_i^\star - f_j^\star| - \max_i |f_i^\lambda - f_i^\star| - \max_j |f_j^\lambda - f_j^\star| \\
&\overset{③}{\geq} \Delta(T^\star) - 0.8X^\star B^\star \gamma/n \\
&\overset{④}{\geq} 2.5009/n - 0.0008/n = 2.5001/n > 2.5/n,
\end{aligned}$$

where ① follows from the definition of the separation distance and ② follows from the triangle inequality. ③ follows from that $\boldsymbol{\theta}^\lambda$ is the fixed point solution of the contraction map (2.19). Thus, $\boldsymbol{\theta}^\lambda \in \mathcal{N}^\star$ following from the non-escaping property by the contraction mapping theorem. This further implies that $\|\mathbf{f}^\lambda - \mathbf{f}^\star\|_\infty \leq 0.4X^\star B^\star\gamma/n$ by (A.10). Finally, ④ follows from that $T^\star$ satisfies the separation condition (2.9): $\Delta(T^\star) \geq 2.5009/n$.

For bounding $\Delta(\tilde{T})$, first identify that $-\max_i |\tilde{f}_i - f_i^\star| \geq -\max_i |f_i^\lambda - f_i^\star|$, since the inner point $\tilde{f}_i$ is included in the interval $[f_i^\star, f_i^\lambda]$ and hence the length of the $[\tilde{f}_i, f_i^\star]$ is less than the entire interval $[f_i^\star, f_i^\lambda]$. Then we immediately arrive at $\Delta(\tilde{T}) > 2.5/n$.

For $\Delta(\hat{T})$, we have

$$\begin{aligned}
\Delta(\hat{T}) &= \min_{i\neq j} |\hat{f}_i - \hat{f}_j| \\
&= \min_{i\neq j} |\hat{f}_i - f_i^\lambda + f_i^\lambda - f_j^\lambda + f_j^\lambda - \hat{f}_j| \\
&\overset{①}{\geq} \min_{i\neq j} |f_i^\lambda - f_j^\lambda| - \max_i |\hat{f}_i - f_i^\lambda| - \max_j |\hat{f}_j - f_j^\lambda| \\
&\overset{②}{\geq} \Delta(T^\lambda) - 2\|\hat{\mathbf{f}} - \mathbf{f}^\lambda\|_\infty \\
&\overset{②}{\geq} \Delta(T^\lambda) - 2(14.08)B^\star\gamma/n,
\end{aligned}$$

where ① follows from the triangle inequality and ② follows from the definition of $\|\hat{\mathbf{f}} - \mathbf{f}^\lambda\|_\infty$. ③ follows from that $\|\hat{\mathbf{f}} - \mathbf{f}^\lambda\|_\infty \leq 0.4(35.2)B^\star\gamma/n = 14.08B^\star\gamma/n$ by (A.10). Finally following from the SNR condition (2.10) and $\Delta(T^\lambda) \geq 2.5001/n$, we then have $\Delta(\hat{T}) \geq 2.5001/n - 2(14.08) \times 10^{-7}/n > 2.5/n$.

$\Delta(\tilde{T}^\lambda) \geq 2.5/n$ holds by the same strategy as $\Delta(\tilde{T}) > 2.5/n$. □

APPENDIX B

APPENDICES FOR CHAPTER 3

B.1 Proof of Lemma 3.4.1

Lemma B.1.1 (Lemma 3.4.1)**.** *The following conditions are necessary for* (3.16)*:*

$$\begin{aligned}
\sum_{j,k} Q_{ijk}\mathbf{v}_p^\star(j)\mathbf{w}_p^\star(k) &= \mathbf{u}_p^\star(i), \forall i \in [n], \forall p \in [r];\\
\sum_{i,k} Q_{ijk}\mathbf{u}_p^\star(i)\mathbf{w}_p^\star(k) &= \mathbf{v}_p^\star(j), \forall i \in [n], \forall p \in [r];\\
\sum_{i,j} Q_{ijk}\mathbf{u}_p^\star(i)\mathbf{v}_p^\star(j) &= \mathbf{w}_p^\star(k), \forall i \in [n], \forall p \in [r]
\end{aligned}$$

or in tensor notation

$$\begin{aligned}
\boldsymbol{\mathcal{Q}}\times_2\mathbf{v}_p^\star\times_3\mathbf{w}_p^\star &= \mathbf{u}_p^\star, \forall p \in [r];\\
\boldsymbol{\mathcal{Q}}\times_1\mathbf{u}_p^\star\times_3\mathbf{w}_p^\star &= \mathbf{v}_p^\star, \forall p \in [r];\\
\boldsymbol{\mathcal{Q}}\times_1\mathbf{u}_p^\star\times_2\mathbf{v}_p^\star &= \mathbf{w}_p^\star, \forall p \in [r]
\end{aligned} \tag{3.21}$$

where $\{\times_k\}$ *are the* k*-mode tensor-vector product [270] whose definitions are apparent from context.*

Proof. From the KKT conditions of the constrained optimization (3.20), we have the partial derivatives of its Lagrangian

$$\mathcal{L}(\mathbf{u},\mathbf{v},\mathbf{w},a,b,c) = q(\mathbf{u},\mathbf{v},\mathbf{w}) - a(\|\mathbf{u}\|_2^2-1) - b(\|\mathbf{v}\|_2^2-1) - c(\|\mathbf{w}\|_2^2-1)$$

at $\mathbf{u}=\mathbf{u}_p^\star$, $\mathbf{v}=\mathbf{v}_p^\star$, and $\mathbf{w}=\mathbf{w}_p^\star$, $p=1,\ldots,r$, must vanish. Therefore,

$$\begin{aligned}
\frac{\partial\mathcal{L}(\mathbf{u}_p^\star,\mathbf{v}_p^\star,\mathbf{w}_p^\star,a,b,c)}{\partial\mathbf{u}} &= \frac{\partial q(\mathbf{u}_p^\star,\mathbf{v}_p^\star,\mathbf{w}_p^\star)}{\partial\mathbf{u}} - 2a\mathbf{u}_p^\star = 0,\\
\frac{\partial\mathcal{L}(\mathbf{u}_p^\star,\mathbf{v}_p^\star,\mathbf{w}_p^\star,a,b,c)}{\partial\mathbf{v}} &= \frac{\partial q(\mathbf{u}_p^\star,\mathbf{v}_p^\star,\mathbf{w}_p^\star)}{\partial\mathbf{v}} - 2b\mathbf{v}_p^\star = 0,\\
\frac{\partial\mathcal{L}(\mathbf{u}_p^\star,\mathbf{v}_p^\star,\mathbf{w}_p^\star,a,b,c)}{\partial\mathbf{w}} &= \frac{\partial q(\mathbf{u}_p^\star,\mathbf{v}_p^\star,\mathbf{w}_p^\star)}{\partial\mathbf{w}} - 2c\mathbf{w}_p^\star = 0.
\end{aligned} \tag{B.1}$$

Hence, $2a = \langle\frac{\partial q(\mathbf{u}_p^\star,\mathbf{v}_p^\star,\mathbf{w}_p^\star)}{\partial\mathbf{u}},\mathbf{u}_p^\star\rangle$, $2b = \langle\frac{\partial q(\mathbf{u}_p^\star,\mathbf{v}_p^\star,\mathbf{w}_p^\star)}{\partial\mathbf{v}},\mathbf{v}_p^\star\rangle$, and $2c = \langle\frac{\partial q(\mathbf{u}_p^\star,\mathbf{v}_p^\star,\mathbf{w}_p^\star)}{\partial\mathbf{u}},\mathbf{w}_p^\star\rangle$. Note that q satisfies the Interpolation condition and $\frac{\partial q(\mathbf{u},\mathbf{v},\mathbf{w})}{\partial\mathbf{u}(i)} = \sum_{j,k} Q_{ijk}\mathbf{v}(j)\mathbf{w}(k)$, we have that

$$2a = \sum_{i,j,k} Q_{ijk}\mathbf{u}_p^\star(i)\mathbf{v}_p^\star(j)\mathbf{w}_p^\star(k) = q(\mathbf{u}_p^\star,\mathbf{v}_p^\star,\mathbf{w}_p^\star) = 1.$$

That is $a = 1/2$. With similar arguments, one can show that $b = c = 1/2$. The conclusion of this lemma follows from (B.1). □

B.2 Proof of Lemma 3.4.2

Lemma B.2.1 (Lemma 3.4.2)**.** *The solution of the least-norm problem (3.22) has the form (normal equation)*

$$\boldsymbol{\mathcal{Q}} = \sum_{p=1}^{r} (\boldsymbol{\alpha}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \boldsymbol{\beta}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \boldsymbol{\gamma}_p^\star) \tag{3.23}$$

with the unknown coefficients $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ *being chosen such that* $\boldsymbol{\mathcal{Q}}$ *in* (3.23) *satisfies* (3.21)*. So we get an explicit form of a pre-certificate*

$$\begin{aligned} q(\mathbf{u}, \mathbf{v}, \mathbf{w}) &= \langle \boldsymbol{\mathcal{Q}}, \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} \rangle \\ &= \sum_{p=1}^{r} [\langle \boldsymbol{\alpha}_p^\star, \mathbf{u} \rangle \langle \mathbf{v}_p^\star, \mathbf{v} \rangle \langle \mathbf{w}_p^\star, \mathbf{w} \rangle + \langle \mathbf{u}_p^\star, \mathbf{u} \rangle \langle \boldsymbol{\beta}_p^\star, \mathbf{v} \rangle \langle \mathbf{w}_p^\star, \mathbf{w} \rangle + \langle \mathbf{u}_p^\star, \mathbf{u} \rangle \langle \mathbf{v}_p^\star, \mathbf{v} \rangle \langle \boldsymbol{\gamma}_p^\star, \mathbf{w} \rangle]. \end{aligned} \tag{3.24}$$

Proof. First, the Lagrangian form of (3.22) is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mathcal{Q}}, \{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r) &= \frac{1}{2}\|\boldsymbol{\mathcal{Q}}\|_F^2 - \sum_{p=1}^{r} \left(\boldsymbol{\mathcal{Q}} \times_1 \boldsymbol{\alpha}_p^\star \times_2 \mathbf{v}_p^\star \times_3 \mathbf{w}_p^\star + \boldsymbol{\mathcal{Q}} \times_1 \mathbf{u}_p^\star \times_2 \boldsymbol{\beta}_p^\star \times_3 \mathbf{w}_p^\star + \boldsymbol{\mathcal{Q}} \times_1 \mathbf{u}_p^\star \times_2 \mathbf{v}_p^\star \times_3 \boldsymbol{\gamma}_p^\star \right) \\ &= \frac{1}{2}\|\boldsymbol{\mathcal{Q}}\|_F^2 - \left\langle \boldsymbol{\mathcal{Q}}, \sum_{p=1}^{r} \boldsymbol{\alpha}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \boldsymbol{\beta}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \boldsymbol{\gamma}_p^\star \right\rangle \end{aligned}$$

with the Lagrangian multipliers $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ to be chosen such that $\boldsymbol{\mathcal{Q}}$ satisfies (3.21). Then, by the KKT necessary conditions, the solution of the least-norm problem (3.22) should satisfy

$$\begin{aligned} \mathbf{0} =& \frac{\partial \mathcal{L}(\boldsymbol{\mathcal{Q}}, \{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r)}{\partial \boldsymbol{\mathcal{Q}}} \\ =& \boldsymbol{\mathcal{Q}} - \sum_{p=1}^{r} \left(\boldsymbol{\alpha}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \boldsymbol{\beta}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \boldsymbol{\gamma}_p^\star \right). \end{aligned}$$

□

B.3 Proof of Lemma 3.4.3

Lemma B.3.1 (Lemma 3.4.3)**.** *Under Assumptions II and III together with* $r = o(n^2/\kappa(\log n)^2)$*, the following estimates are valid for sufficiently large* n*:*

$$\begin{aligned} \left\| \mathbf{A} - \frac{1}{3}\mathbf{U} \right\| &\leq 2\kappa(\log n) \left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}} \right); \\ \left\| \mathbf{B} - \frac{1}{3}\mathbf{V} \right\| &\leq 2\kappa(\log n) \left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}} \right); \\ \left\| \mathbf{C} - \frac{1}{3}\mathbf{W} \right\| &\leq 2\kappa(\log n) \left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}} \right) \end{aligned}$$

where

$$\mathbf{A} = \left[\boldsymbol{\alpha}_1^\star, \cdots, \boldsymbol{\alpha}_r^\star\right], \mathbf{U} = \left[\mathbf{u}_1^\star, \cdots, \mathbf{u}_r^\star\right];$$
$$\mathbf{B} = \left[\boldsymbol{\beta}_1^\star, \cdots, \boldsymbol{\beta}_r^\star\right], \mathbf{V} = \left[\mathbf{v}_1^\star, \cdots, \mathbf{v}_r^\star\right];$$
$$\mathbf{C} = \left[\boldsymbol{\gamma}_1^\star, \cdots, \boldsymbol{\gamma}_r^\star\right], \mathbf{W} = \left[\mathbf{w}_1^\star, \cdots, \mathbf{w}_r^\star\right]$$

and the norm $\|\cdot\|$ is the matrix spectral norm.

Proof. We need to find coefficient vectors $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ so that the tensor

$$\boldsymbol{\mathcal{Q}} = \sum_{p=1}^r \left(\boldsymbol{\alpha}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \boldsymbol{\beta}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \boldsymbol{\gamma}_p^\star\right)$$

satisfies (3.21):

$$\begin{aligned}
&\boldsymbol{\mathcal{Q}} \times_2 \mathbf{v}_p^\star \times_3 \mathbf{w}_p^\star = \mathbf{u}_p^\star, \quad \forall p \in [r],\\
&\boldsymbol{\mathcal{Q}} \times_1 \mathbf{u}_p^\star \times_3 \mathbf{w}_p^\star = \mathbf{v}_p^\star, \quad \forall p \in [r],\\
&\boldsymbol{\mathcal{Q}} \times_1 \mathbf{u}_p^\star \times_2 \mathbf{v}_p^\star = \mathbf{w}_p^\star, \quad \forall p \in [r].
\end{aligned} \tag{B.2}$$

An iteration scheme. We adopt the following *iterative scheme* to find such $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$:

$$\begin{aligned}
\boldsymbol{\alpha}_q^{t+1} &= \boldsymbol{\alpha}_q^t - \rho\left(\boldsymbol{\mathcal{Q}}_1^t \times_2 \mathbf{v}_p^\star \times_3 \mathbf{w}_q^\star - \mathbf{u}_q^\star\right), \quad q \in [r],\\
\boldsymbol{\beta}_q^{t+1} &= \boldsymbol{\beta}_q^t - \rho\left(\boldsymbol{\mathcal{Q}}_2^t \times_1 \mathbf{u}_p^\star \times_3 \mathbf{w}_q^\star - \mathbf{v}_q^\star\right), \quad q \in [r],\\
\boldsymbol{\gamma}_q^{t+1} &= \boldsymbol{\gamma}_q^t - \rho\left(\boldsymbol{\mathcal{Q}}_3^t \times_1 \mathbf{u}_p^\star \times_2 \mathbf{v}_q^\star - \mathbf{w}_q^\star\right), \quad q \in [r],
\end{aligned} \tag{B.3}$$

initialized by $\boldsymbol{\alpha}_q^0 = \frac{1}{3}\mathbf{u}_q^\star$, $\boldsymbol{\beta}_q^0 = \frac{1}{3}\mathbf{v}_q^\star$, and $\boldsymbol{\gamma}_q^0 = \frac{1}{3}\mathbf{w}_q^\star$ with $q \in [r]$. Here the parameter ρ is a step size to be chosen later and the tensors

$$\begin{aligned}
\boldsymbol{\mathcal{Q}}_1^t &:= \sum_{p=1}^r \left(\boldsymbol{\alpha}_p^t \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \boldsymbol{\beta}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \boldsymbol{\gamma}_p^\star\right),\\
\boldsymbol{\mathcal{Q}}_2^t &:= \sum_{p=1}^r \left(\boldsymbol{\alpha}_p^t \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \boldsymbol{\beta}_p^t \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \boldsymbol{\gamma}_p^\star\right),\\
\boldsymbol{\mathcal{Q}}_3^t &:= \sum_{p=1}^r \left(\boldsymbol{\alpha}_p^t \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \boldsymbol{\beta}_p^t \otimes \mathbf{w}_p^\star + \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \boldsymbol{\gamma}_p^t\right).
\end{aligned} \tag{B.4}$$

Note that the above iterative scheme is for theoretical analysis only as we used $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ in the definitions of $\boldsymbol{\mathcal{Q}}_1^t$, $\boldsymbol{\mathcal{Q}}_2^t$ and $\boldsymbol{\mathcal{Q}}_3^t$.

Convergence of the iteration scheme. We next establish the convergence of the iterations (B.3). Plugging the tensor eigenvalue equations (B.2) into (B.3) followed by subtracting the true solutions from both sides yields for $q \in [r]$

$$
\begin{aligned}
\boldsymbol{\alpha}_q^{t+1} - \boldsymbol{\alpha}_q^\star =& \boldsymbol{\alpha}_q^t - \boldsymbol{\alpha}_q^\star - \rho[\boldsymbol{\mathcal{Q}}_1^t - \boldsymbol{\mathcal{Q}}]\times_2 \mathbf{v}_q^\star \times_3 \mathbf{w}_q^\star, \\
\boldsymbol{\beta}_q^{t+1} - \boldsymbol{\beta}_q^\star =& \boldsymbol{\beta}_q^t - \boldsymbol{\beta}_q^\star - \rho[\boldsymbol{\mathcal{Q}}_2^t - \boldsymbol{\mathcal{Q}}]\times_1 \mathbf{u}_q^\star \times_3 \mathbf{w}_q^\star, \\
\boldsymbol{\gamma}_q^{t+1} - \boldsymbol{\gamma}_q^\star =& \boldsymbol{\gamma}_q^t - \boldsymbol{\gamma}_q^\star - \rho[\boldsymbol{\mathcal{Q}}_3^t - \boldsymbol{\mathcal{Q}}]\times_1 \mathbf{u}_q^\star \times_2 \mathbf{v}_q^\star.
\end{aligned} \tag{B.5}
$$

Then plugging the definitions of $\boldsymbol{\mathcal{Q}}_1^t, \boldsymbol{\mathcal{Q}}_2^t, \boldsymbol{\mathcal{Q}}_3^t$ (B.4) into (B.5) and using the following matrix notations

$$
\begin{aligned}
A^t &:= \left[\boldsymbol{\alpha}_1^t, \cdots, \boldsymbol{\alpha}_r^t\right], \mathbf{A} := \left[\boldsymbol{\alpha}_1^\star, \cdots, \boldsymbol{\alpha}_r^\star\right], \\
\mathbf{B}^t &:= \left[\boldsymbol{\alpha}_1^t, \cdots, \boldsymbol{\alpha}_r^t\right], \mathbf{B} := \left[\boldsymbol{\alpha}_1^\star, \cdots, \boldsymbol{\alpha}_r^\star\right], \\
\mathbf{C}^t &:= \left[\boldsymbol{\gamma}_1^t, \cdots, \boldsymbol{\gamma}_r^t\right], \mathbf{C} := \left[\boldsymbol{\gamma}_1^\star, \cdots, \boldsymbol{\gamma}_r^\star\right],
\end{aligned}
$$

we have

$$
\begin{aligned}
\mathbf{A}^{t+1} - \mathbf{A} =& (\mathbf{A}^t - \mathbf{A})(\mathbf{I} - \rho\left[(\mathbf{V}^\top\mathbf{V}) \odot (\mathbf{W}^\top\mathbf{W})\right]), \\
\mathbf{B}^{t+1} - \mathbf{B} =& (\mathbf{B}^t - \mathbf{B})\big(\mathbf{I} - \rho[(\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{W}^\top\mathbf{W})]\big) - \rho\mathbf{V}\left[((\mathbf{A}^t - \mathbf{A})^\top\mathbf{U}) \odot (\mathbf{W}^\top\mathbf{W})\right], \\
\mathbf{C}^{t+1} - \mathbf{C} =& (\mathbf{C}^t - \mathbf{C})(\mathbf{I} - \rho[(\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V})]) - \rho\mathbf{W}\left\{[((\mathbf{A}^t - \mathbf{A})^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V})] + [(\mathbf{U}^\top\mathbf{U}) \odot ((\mathbf{B}^t - \mathbf{B})^\top\mathbf{V})]\right\}.
\end{aligned} \tag{B.6}
$$

Denoting $e_a^t = \|\mathbf{A}^t - \mathbf{A}\|, e_b^t = \|\mathbf{B}^t - \mathbf{B}\|, e_c^t = \|\mathbf{C}^t - \mathbf{C}\|$ and

$$
\tilde{\rho} := \rho \min\left\{\begin{matrix} \lambda_{\min}((\mathbf{V}^\top\mathbf{V}) \odot (\mathbf{W}^\top\mathbf{W})) \\ \lambda_{\min}((\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{W}^\top\mathbf{W})) \\ \lambda_{\min}((\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V})) \end{matrix}\right\},
$$

it follows from (B.6) that

$$
\begin{aligned}
e_a^{t+1} &\leq (1 - \tilde{\rho})e_a^t, \\
e_b^{t+1} &\leq \rho\|\mathbf{U}\|\|\mathbf{V}\|\|\mathbf{W}\|^2 e_a^t + (1 - \tilde{\rho})e_b^t, \\
e_c^{t+1} &\leq \rho\|\mathbf{U}\|^2\|\mathbf{V}\|\|\mathbf{W}\|e_a^t + \rho\|\mathbf{U}\|^2\|\mathbf{V}\|\|\mathbf{W}\|e_b^t + (1 - \tilde{\rho})e_c^t,
\end{aligned} \tag{B.7}
$$

where we have used triangle inequality and properties of spectral norms such as $\|\mathbf{P} \odot \mathbf{Q}\| \leq \|\mathbf{P}\|\|\mathbf{Q}\|$ [43]. Converting (B.7) into matrix form gives

$$
\begin{bmatrix} e_a^{t+1} \\ e_b^{t+1} \\ e_c^{t+1} \end{bmatrix} \leq \begin{bmatrix} 1 - \tilde{\rho} & 0 & 0 \\ \rho\|\mathbf{U}\|\|\mathbf{V}\|\|\mathbf{W}\|^2 & 1 - \tilde{\rho} & 0 \\ \rho\|\mathbf{U}\|\|\mathbf{W}\|\|\mathbf{V}\|^2 & \rho\|\mathbf{U}\|^2\|\mathbf{V}\|\|\mathbf{W}\| & 1 - \tilde{\rho} \end{bmatrix} \begin{bmatrix} e_a^t \\ e_b^t \\ e_c^t \end{bmatrix},
$$

where the lower triangular system matrix share the same value

[43] Hadamard product $\mathbf{P} \odot \mathbf{Q}$ is a principal submatrix of $\mathbf{P} \otimes \mathbf{Q}$, whose singular values are the products of the individual singular values of $\mathbf{P}$ and $\mathbf{Q}$.

$$
\begin{aligned}
\eta &= 1-\tilde{\rho} \\
&\in \left[1-\rho\left(1+\frac{\kappa(\log n)\sqrt{r}}{n}\right), 1-\rho\left(1-\frac{\kappa(\log n)\sqrt{r}}{n}\right)\right] \qquad \text{(B.8)} \\
&\subset (0,1)
\end{aligned}
$$

where (B.8) follows from applying Weyl's inequality to (3.6) in Assumption III and the last line holds for any $\rho \in \left(0, (1+\frac{\kappa(\log n)\sqrt{r}}{n})^{-1}\right)$.

Conclusion. The error sequence (e_a^t, e_b^t, e_c^t) is convergent to $(0,0,0)$ geometrically with a rate $\eta \in (0,1)$. Thus,

$$\lim_{t\to\infty}(\mathbf{A}^t, \mathbf{B}^t, \mathbf{C}^t) = (\mathbf{A}, \mathbf{B}, \mathbf{C}).$$

Convergence of the consecutive differences $\{\|\mathbf{A}^t - \mathbf{A}^{t-1}\|\}, \{\|\mathbf{B}^t - \mathbf{B}^{t-1}\|\}, \{\|\mathbf{C}^t - \mathbf{C}^{t-1}\|\}$. Subtracting the following two consecutive iterations for $\{\mathbf{A}^t\}$ in (B.6):

$$
\begin{aligned}
\mathbf{A}^{t+1} - \mathbf{A} &= (\mathbf{A}^t - \mathbf{A})(\mathbf{I} - \rho\left[(\mathbf{V}^\top\mathbf{V}) \odot (\mathbf{W}^\top\mathbf{W})\right]) \\
\mathbf{A}^t - \mathbf{A} &= (\mathbf{A}^{t-1} - \mathbf{A})(\mathbf{I} - \rho\left[(\mathbf{V}^\top\mathbf{V}) \odot (\mathbf{W}^\top\mathbf{W})\right])
\end{aligned}
$$

yields

$$\mathbf{A}^{t+1} - \mathbf{A}^t = (\mathbf{A}^t - A^{t-1})(\mathbf{I} - \rho\left[(\mathbf{V}^\top\mathbf{V}) \odot (\mathbf{W}^\top\mathbf{W})\right]).$$

Similar manipulations applied to $\{\mathbf{B}^t\}$ and $\{\mathbf{C}^t\}$ lead to

$$
\begin{aligned}
\mathbf{B}^{t+1} - \mathbf{B}^t =& (\mathbf{B}^t - \mathbf{B}^{t-1})(\mathbf{I} - \rho\left[(\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{W}^\top\mathbf{W})\right]) - \rho\mathbf{V}\left[((\mathbf{A}^t - \mathbf{A}^{t-1})^\top\mathbf{U}) \odot (\mathbf{W}^\top\mathbf{W})\right], \\
\mathbf{C}^{t+1} - \mathbf{C}^t =& (\mathbf{C}^t - \mathbf{C}^{t-1})(\mathbf{I} - \rho\left[(\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V})\right]) \\
& - \rho\mathbf{W}\left\{\left[((\mathbf{A}^t - \mathbf{A}^{t-1})^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V})\right] + \left[(\mathbf{U}^\top\mathbf{U}) \odot ((\mathbf{B}^t - \mathbf{B}^{t-1})^\top\mathbf{V})\right]\right\}
\end{aligned}
$$

Defining $\hat{e}_a^t = \|\mathbf{A}^t - \mathbf{A}^{t-1}\|, \hat{e}_b^t = \|\mathbf{B}^t - \mathbf{B}^{t-1}\|, \hat{e}_c^t = \|\mathbf{C}^t - \mathbf{C}^{t-1}\|$, we can get the same form as (B.7) and therefore claim that $(\hat{e}_a^t, \hat{e}_b^t, \hat{e}_c^t)$ converge to $(0,0,0)$ geometrically with the same rate $\eta \in (0,1)$ in (B.8).

Bounding the accumulative errors. The geometric convergence of $\{\|\mathbf{C}^t - \mathbf{C}^{t-1}\|\}$ implies

$$\|\mathbf{C}^t - \mathbf{C}^{t-1}\| \le \eta^{t-1}\|\mathbf{C}^1 - \mathbf{C}^0\|$$

which together with the triangle inequality gives

$$\|\mathbf{C}^t - \mathbf{C}^0\| \le \sum_{s=0}^{t-1}\|\mathbf{C}^{s+1} - \mathbf{C}^s\| \le \sum_{s=0}^{t-1}\eta^s\|\mathbf{C}^1 - \mathbf{C}^0\| \le \frac{1}{1-\eta}\|\mathbf{C}^1 - \mathbf{C}^0\|.$$

Letting $\mathcal{T}$ go to infinity on the left-hand side gives

$$\|\mathbf{C}-\mathbf{C}^0\| \leq \frac{1}{1-\eta}\|\mathbf{C}^1-\mathbf{C}^0\|. \tag{B.9}$$

We next bound $\|\mathbf{C}^1-\mathbf{C}^0\|$. From (B.3), we have

$$\boldsymbol{\gamma}_q^1-\boldsymbol{\gamma}_q^0=\rho(\boldsymbol{\mathcal{Q}}_3^0\times_1\mathbf{u}_q^\star\times_2\mathbf{v}_q^\star-\mathbf{w}_q^\star)=\rho\left(\sum_{p=1}^r\langle\mathbf{u}_p^\star,\mathbf{u}_q^\star\rangle\langle\mathbf{v}_p^\star,\mathbf{v}_q^\star\rangle\mathbf{w}_p^\star-\mathbf{w}_q^\star\right)$$

implying

$$\mathbf{C}^1-\mathbf{C}^0=\rho\mathbf{W}((\mathbf{U}^\top\mathbf{U})\odot(\mathbf{V}^\top\mathbf{V})-\mathbf{I}).$$

Then from Assumptions II and III, we have

$$\|\mathbf{C}^1-\mathbf{C}^0\|\leq\rho\|\mathbf{W}\|\|(\mathbf{U}^\top\mathbf{U})\odot(\mathbf{V}^\top\mathbf{V})-\mathbf{I}\|\leq\rho\left(1+c\sqrt{\frac{r}{n}}\right)\frac{\kappa(\log n)\sqrt{r}}{n}. \tag{B.10}$$

Combine ALL. Finally, combining (B.8), (B.9) and (B.10) and using $\mathbf{C}_0=\frac{1}{3}\mathbf{W}$, we have

$$\begin{aligned}\left\|\mathbf{C}-\frac{1}{3}\mathbf{W}\right\|\leq&\frac{1+c\sqrt{\frac{r}{n}}}{1-\frac{\kappa(\log n)\sqrt{r}}{n}}\frac{\kappa(\log n)\sqrt{r}}{n}\\ \leq&2\left(1+c\sqrt{\frac{r}{n}}\right)\frac{\kappa(\log n)\sqrt{r}}{n}\\ =&2\kappa(\log n)\left(\frac{\sqrt{r}}{n}+c\frac{r}{n^{1.5}}\right)\end{aligned}$$

where the second line follows from the assumption $r=o(n^2/\kappa(\log n)^2)$ which implies $1-\frac{\kappa(\log n)\sqrt{r}}{n}\geq\frac{1}{2}$ for a sufficiently large n. Similar arguments and bounds apply to $\|\mathbf{A}-\frac{1}{3}\mathbf{U}\|$ and $\|\mathbf{B}-\frac{1}{3}\mathbf{V}\|$.

□

B.4 Proof of Lemma 3.4.4

Lemma B.4.1 (Lemma 3.4.4)**.** *Under Assumptions I, II, III, if $r\ll n^{1.25}$ and $r\leq\frac{n}{24\delta c^2}$ for $\delta\in(0,\frac{1}{24}]$, then for sufficiently large n, we have $|q(\mathbf{u},\mathbf{v},\mathbf{w})|<1$ in $\mathcal{F}(\delta)$.*

Proof. The following lemma is required in the proof of Lemma3.4.4. Let us first admit Lemma B.4.2 to prove Lemma3.4.4. Since q is the sum of two parts given in (3.29) and (3.30), to bound $|q|$, we will control these parts separately.

Lemma B.4.2. *Under Assumptions I and II, if $r\leq n^{1.25-1.5r_c}$ with $r_c\in(0,1/6)$, then for any integer $p\geq3$,*

$$\|\mathbf{U}^\top\|_{2\to p} \le 1 + \frac{1}{p}\tau(\log n)n^{-r_c}$$

The same bounds hold for $\mathbf{V}$ *and* $\mathbf{W}$. *Here, we define* $\|\mathbf{H}\|_{2\to p} := \sup\{\|\mathbf{H}\mathbf{x}\|_p : \mathbf{x} \in \mathbb{S}^{n-1}\}$.

Proof of Lemma B.4.2. See Section B.4.1. □

Bound absolute value of (3.29)**.**

$$\begin{aligned}
\sum_{p=1}^{r} |\langle \boldsymbol{\alpha}_p^\star - \frac{1}{3}\mathbf{u}_p^\star, \mathbf{u}\rangle\langle \mathbf{v}_p^\star, \mathbf{v}\rangle\langle \mathbf{w}_p^\star, \mathbf{w}\rangle| &\le \sqrt{\sum_{p=1}^{r}\langle \boldsymbol{\alpha}_p^\star - \frac{1}{3}\mathbf{u}_p^\star, \mathbf{u}\rangle^2}\sqrt{\sum_{p=1}^{r}\langle \mathbf{v}_p^\star, \mathbf{v}\rangle^2\langle \mathbf{w}_p^\star, \mathbf{w}\rangle^2} \\
&\le \sqrt{\sum_{p=1}^{r}\langle \boldsymbol{\alpha}_p^\star - \frac{1}{3}\mathbf{u}_p^\star, \mathbf{u}\rangle^2}\sqrt{\sum_{p=1}^{r}\langle \mathbf{v}_p^\star, \mathbf{v}\rangle^4}\sqrt{\sum_{p=1}^{r}\langle \mathbf{w}_p^\star, \mathbf{w}\rangle^4} \\
&= \|(\mathbf{A} - \frac{1}{3}\mathbf{U})^\top\mathbf{u}\|_2\|\mathbf{V}^\top\mathbf{v}\|_4\|\mathbf{W}^\top\mathbf{w}\|_4 \\
&\le \|\mathbf{A} - \frac{1}{3}\mathbf{U}\|\|\mathbf{V}^\top\|_{2\to 4}\|\mathbf{W}^\top\|_{2\to 4} \\
&\le 2\kappa(\log n)\left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}}\right)(1 + o(1)) \\
&= o(1),
\end{aligned}$$

where the last second line follows from Lemma3.4.3 and LemmaB.4.2 when $r \ll n^{1.25}$ (by letting r_c in "$r \ll n^{1.25-r_c}$" approach to zero). The last line holds for $r \ll \frac{n^{1.5}}{\kappa(\log n)}$.

Similar bounds hold for the other two terms in (3.29).

Bound the absolute value of (3.30)**.** First of all, for any $(\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathcal{F}(\delta)$, there exists a division of $[r] = \Omega_u \cup \Omega_v \cup \Omega_w$ such that

$$\begin{aligned}
|\langle \mathbf{u}_p^\star, \mathbf{u}\rangle| &\le \delta, \quad \forall p \in \Omega_u, \\
|\langle \mathbf{v}_p^\star, \mathbf{v}\rangle| &\le \delta, \quad \forall p \in \Omega_v, \\
|\langle \mathbf{w}_p^\star, \mathbf{u}\rangle| &\le \delta, \quad \forall p \in \Omega_w.
\end{aligned} \tag{B.11}$$

We will denote by $\mathbf{U}_{\Omega_u}$ the submatrix of $\mathbf{U}$ forming from those columns of $\mathbf{U}$ with indexes in Ω_u. Similarly, we can define $\mathbf{V}_{\Omega_v}$ and $\mathbf{W}_{\Omega_w}$. With these preparation, we have that

$$\begin{aligned}\sum_{p=1}^{r}|\langle\mathbf{u}_p^\star,\mathbf{u}\rangle\langle\mathbf{v}_p^\star,\mathbf{v}\rangle\langle\mathbf{w}_p^\star\mathbf{w}\rangle| &= \sum_{p\in\Omega_u\cup\Omega_v\cup\Omega_w}|\langle\mathbf{u}_p^\star,\mathbf{u}\rangle\langle\mathbf{v}_p^\star,\mathbf{v}\rangle\langle\mathbf{w}_p^\star\mathbf{w}\rangle| \\ &\le \delta(\|\mathbf{V}_{\Omega_u}\|\|\mathbf{W}_{\Omega_u}\|+\|\mathbf{U}_{\Omega_v}\|\|\mathbf{W}_{\Omega_v}\|+\|\mathbf{U}_{\Omega_w}\|\|\mathbf{V}_{\Omega_w}\|) \\ &\le 3\delta\left(1+c\sqrt{\frac{r}{n}}\right)^2 \\ &\le 12\delta\max\{1,c^2r/n\} \\ &\le \frac{1}{2},\end{aligned}$$

where the first inequality follows from (B.11) and $\sum_{p\in\Omega_u}|\langle\mathbf{v}_p^\star,\mathbf{v}\rangle\langle\mathbf{w}_p^\star,\mathbf{w}\rangle| \le \|\mathbf{V}_{\Omega_u}\|\|\mathbf{W}_{\Omega_u}\|$, etc. The second inequality uses the fact that the spectral norm of any submatrix is smaller than the original one and Assumption II. The last inequality holds when $\delta \le \frac{1}{24}$ and and $r \le n/(24\delta c^2)$.

Combine ALL. Under Assumptions I, II, III, if $r \ll n^{1.25}$ and $r \le \frac{n}{24\delta c^2}$ for $\delta \in (0, \frac{1}{24}]$, we have $|q| \le o(1) + \frac{1}{2} < 1$ in $\mathcal{F}(\delta)$ for sufficiently large n.

□

B.4.1 Proof of Lemma B.4.2

The proof refines the one for Lemma 4 of [71]. We only prove it for $\mathbf{U}$ since the same arguments apply to $\mathbf{W}$ and $\mathbf{V}$. We start with a general integer $p \ge 3$.

$$\|\mathbf{U}^\top\|_{2\to p} = \sup_{\mathbf{x}\in\mathbb{S}^{n-1}}\|\mathbf{U}^\top\mathbf{x}\|_p := \|\mathbf{U}^\top\mathbf{x}^\star\|_p \tag{B.12}$$

where we define $\mathbf{x}^\star \in \mathbb{S}^{n-1}$ to be the optimal solution of $\sup_{\mathbf{x}\in\mathbb{S}^{n-1}}\|\mathbf{U}^\top\mathbf{x}\|_p^p$. Further note that

$$\|\mathbf{U}^\top\mathbf{x}^\star\|_p^p = \|\mathbf{U}_S^\top\mathbf{x}^\star\|_p^p + \|\mathbf{U}_{S^c}^\top\mathbf{x}^\star\|_p^p \tag{B.13}$$

where S denotes the indices of the largest (in absolute value) L entries of $\mathbf{U}^\top\mathbf{x}^\star$ and $\mathbf{U}_S$ denotes the column submatrix of $\mathbf{U}$ indexed by S. Similar notations apply to its complement set $S^c = [r] \setminus S$.

Bound the first term.

$$\|\mathbf{U}_S^\top\mathbf{x}^\star\|_p^p \le \|\mathbf{U}_S^\top\mathbf{x}^\star\|_2^2 \le \|\mathbf{U}_S\mathbf{U}_S^\top\| \le 1 + \sum_{i\in S\setminus\{j\}}|\langle\mathbf{u}_i,\mathbf{u}_j\rangle| \le 1 + (L-1)\frac{\tau(\log n)}{\sqrt{n}}. \tag{B.14}$$

Note this upper-bound is independent of p. Here, the first inequality is because $|\mathbf{u}_i^{\star\top}\mathbf{x}^\star| \le \|\mathbf{u}_i^\star\|_2\|\mathbf{x}^\star\|_2 = 1$ and the last second inequality follows from *Gershgorin's circle theorem*. Finally the last inequality is from Assumption I and L being the cardinality of the set S.

Bound the second term. First note that

$$\min_{i\in S}|\mathbf{u}_i^\top \mathbf{x}^\star|^2 \le \frac{1}{L}\sum_{i\in S}|\mathbf{u}_i^\top \mathbf{x}^\star|^2 \le \frac{1}{L}\|\mathbf{U}_S\mathbf{U}_S^\top\|\|\mathbf{x}^\star\|_2^2 \le \frac{1}{L}(1+o(1)) \le \frac{2}{L}$$

for sufficiently large n. The last second inequality follows from (B.14) and an additional assumption on L

$$(L-1)\frac{\tau(\log n)}{\sqrt{n}} = o(1). \tag{B.15}$$

We conclude that

$$\max_{i\in S^c}|\mathbf{u}_i^\top \mathbf{x}^\star|^2 \le \min_{i\in S}|\mathbf{u}_i^\top \mathbf{x}^\star|^2 \le \frac{2}{L},$$

since S consists of the indices of the L largest (in absolute value) elements of $\mathbf{U}^\top\mathbf{x}^\star$. As a consequence, we have

$$\begin{aligned}\|\mathbf{U}_{S^c}^\top \mathbf{x}^\star\|_p^p = \sum_{i\notin S}|\mathbf{u}_i^\top \mathbf{x}^\star|^p \le \big(\max_{i\notin S}|\mathbf{u}_i^\top \mathbf{x}^\star|^{p-2}\big)\sum_{i\notin S}|\mathbf{u}_i^\top \mathbf{x}|^2 = \big(\max_{i\notin S}|\mathbf{u}_i^\top \mathbf{x}^\star|^{p-2}\big)\|\mathbf{U}_{S^c}^\top x^\star\|_2^2 \\ \le \left(\frac{2}{L}\right)^{\frac{p}{2}-1}\left(1+c\sqrt{\frac{r}{n}}\right)^2 \end{aligned} \tag{B.16}$$

where the last inequality follows from the fact that $\|\mathbf{U}_{S^c}^\top \mathbf{x}^\star\|_2^2 \le \|\mathbf{U}_{S^c}\|^2 \le \|\mathbf{U}\|^2 \le (1+c\sqrt{\frac{r}{n}})^2$ by Assumption II. Furthermore, since $(1+c\sqrt{\frac{r}{n}})^2 \le 4\max\{1, c^2\frac{r}{n}\}$, $c^2\frac{r}{n} \le c^2 n^{0.25-1.5r_c}$ from the condition of $r \le n^{1.25-1.5r_c}$, and $1 \ll c^2 n^{0.25-1.5r_c}$ for $r_c \in (0, 1/6)$, we have $(1+c\sqrt{\frac{r}{n}})^2 \le 4c^2 n^{0.25-1.5r_c}$ for $r_c \in (0, 1/6)$. So from (B.16), we get

$$\|\mathbf{U}_{S^c}^\top \mathbf{x}^\star\|_p^p \le 4\left(\frac{2}{L}\right)^{\frac{p}{2}-1} c^2 n^{0.25-1.5r_c}. \tag{B.17}$$

From (B.13), (B.14), and (B.17), we have

$$\|\mathbf{U}^\top \mathbf{x}^\star\|_p^p \le 1 + (L-1)\frac{\tau(\log n)}{\sqrt{n}} + 4\left(\frac{2}{L}\right)^{\frac{p}{2}-1} c^2 n^{0.25-1.5r_c}.$$

By choosing

$$L = \left\lceil \frac{1}{2}n^{0.5-r_c}\right\rceil \Rightarrow \begin{cases} L \le \frac{1}{2}n^{0.5-r_c}+1 \\ L \ge \frac{1}{2}n^{0.5-r_c}\end{cases}$$

which satisfies the condition (B.15), we have that

$$\|\mathbf{U}^\top \mathbf{x}^\star\|_p^p \le 1 + \frac{1}{2}\tau(\log n)n^{-r_c} + 4^{\frac{p}{2}}c^2 n^{(\frac{3}{4}-\frac{p}{4})+(\frac{p}{2}-\frac{5}{2})r_c}.$$

Then from the assumptions $p \ge 3$ and $r_c \in (0, \frac{1}{6})$, we get

$$\left(\frac{3}{4}-\frac{p}{4}\right)+\left(\frac{p}{2}-\frac{5}{2}\right)r_c \le \left(\frac{3}{4}-\frac{p}{4}\right)6r_c+\left(\frac{p}{2}-\frac{5}{2}\right)r_c = (2-p)r_c \le -r_c. \tag{B.18}$$

So, we have

$$\|\mathbf{U}^\top \mathbf{x}^\star\|_p^p \le 1+\left(\frac{1}{2}\tau(\log n)+4^{\frac{p}{2}}c^2\right)n^{-r_c}.$$

Since $4^{\frac{p}{2}}c^2 \ll \frac{1}{2}\tau(\log n)$ and $(1+t)^{1/p} \le 1+\frac{1}{p}t$ for all $t \ge 0$, then

$$\|\mathbf{U}^\top \mathbf{x}^\star\|_p \le 1+\frac{1}{p}\tau(\log n)n^{-r_c}$$

holds for any $p \ge 3$. This completes the proof since $\|\mathbf{U}^\top\|_{2\to p} = \|\mathbf{U}^\top \mathbf{x}^\star\|_p$ by (B.12).

B.5 Proof of Lemma 3.4.5

Lemma B.5.1 (Lemma 3.4.5)**.** *Under Assumptions I, II, III, if $r \le n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$, then for sufficiently large n, we have*

$$|F(\theta_1,\theta_2,\theta_3)| \le |\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)| + |\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)| + \frac{4}{3}\tau(\log n)n^{-r_c}. \tag{3.38}$$

Proof. We start by the angular dual polynomial (3.37)

$$\begin{aligned} q(\mathbf{u}(\theta_1),\mathbf{v}(\theta_2),\mathbf{w}(\theta_3)) = \cos(\theta_1)\cos(\theta_2)\cos(\theta_3) &+ q(\mathbf{u}_1^\star,\mathbf{y},\mathbf{z})\cos(\theta_1)\sin(\theta_2)\sin(\theta_3) \\ &+ q(\mathbf{x},\mathbf{v}_1^\star,\mathbf{z})\sin(\theta_1)\cos(\theta_2)\sin(\theta_3) \\ &+ q(\mathbf{x},\mathbf{y},\mathbf{w}_1^\star)\sin(\theta_1)\sin(\theta_2)\cos(\theta_3) \\ &+ q(\mathbf{x},\mathbf{y},\mathbf{z})\sin(\theta_1)\sin(\theta_2)\sin(\theta_3). \end{aligned}$$

To bound q, we only need to bound the coefficients $q(\mathbf{u}_1^\star,\mathbf{y},\mathbf{z})$, $q(\mathbf{x},\mathbf{v}_1^\star,\mathbf{z})$, $q(\mathbf{x},\mathbf{y},\mathbf{w}_1^\star)$, and $q(\mathbf{x},\mathbf{y},\mathbf{z})$.

We first show that $q(\mathbf{u}_1^\star,\mathbf{y},\mathbf{z})$, $q(\mathbf{x},\mathbf{v}_1^\star,\mathbf{z})$, and $q(\mathbf{x},\mathbf{y},\mathbf{w}_1^\star)$ are close to zero. To see this, we examine

$$\begin{aligned} q(\mathbf{x},\mathbf{y},\mathbf{w}_1^\star) =& \sum_{p=1}^{r}[\langle\boldsymbol{\alpha}_p^\star,\mathbf{x}\rangle\langle\mathbf{v}_p^\star,\mathbf{y}\rangle\langle\mathbf{w}_p^\star,\mathbf{w}_1^\star\rangle + \langle\mathbf{u}_p^\star,\mathbf{x}\rangle\langle\boldsymbol{\beta}_p^\star,\mathbf{y}\rangle\langle\mathbf{w}_p^\star,\mathbf{w}_1^\star\rangle \\ &+ \langle\mathbf{u}_p^\star,\mathbf{x}\rangle\langle\mathbf{v}_p^\star,\mathbf{y}\rangle\langle\boldsymbol{\gamma}_p^\star,\mathbf{w}_1^\star\rangle] \\ =& \mathbf{x}^\top[\mathbf{A}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{V}^\top + \mathbf{U}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{B}^\top + \mathbf{U}\,\mathrm{diag}(\mathbf{C}^\top\mathbf{w}_1^\star)\mathbf{V}^\top]\mathbf{y} \\ =& \mathbf{x}^\top\left(\mathbf{A}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^\star + \mathbf{U}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{B}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^\star + \mathbf{U}\,\mathrm{diag}(\mathbf{C}^\top\mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^\star\right)\mathbf{y}, \end{aligned}$$

since $\mathbf{x} \perp \mathbf{u}_1^\star, \mathbf{y} \perp \mathbf{v}_1^\star$. This implies

$$|q(\mathbf{x},\mathbf{y},\mathbf{w}_1^\star)| \le \left\|\mathbf{A}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^\star\right\| + \left\|\mathbf{U}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{B}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^\star\right\| + \left|\mathbf{x}^\top\left(\mathbf{U}\,\mathrm{diag}(\mathbf{C}^\top\mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^\star\right)\mathbf{y}\right|.$$

We first bound $\left\|\mathbf{A}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^\star\right\|$.

$$\begin{aligned}\left\|\mathbf{A}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^{\star\top}\right\| \le& \left\|\mathbf{A}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{U}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{V}^\top\right\| \\ &+ \left\|\frac{1}{3}\mathbf{U}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^{\star\top}\right\| \\ \le& \left\|\mathbf{A} - \frac{1}{3}\mathbf{U}\right\|\|\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\|\|\mathbf{V}\| + \frac{1}{3}\|\mathbf{U}\|\|\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star - \mathbf{e}_1)\|\mathbf{V}^\top\| \\ \le& 2\kappa(\log n)\left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}}\right)\left(1 + c\sqrt{\frac{r}{n}}\right) + \frac{\tau(\log n)}{3\sqrt{n}}\left(1 + c\sqrt{\frac{r}{n}}\right)^2 \\ =& \left[2\kappa(\log n)\frac{\sqrt{r}}{n} + \frac{\tau(\log n)}{3\sqrt{n}}\right]\left(1 + c\sqrt{\frac{r}{n}}\right)^2,\end{aligned} \tag{B.19}$$

where the third inequality first uses the facts $\|\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\| = 1$ and $\|\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star - \mathbf{e}_1)\| = \max_{p\neq 1}|\langle\mathbf{w}_p^\star,\mathbf{w}_1^\star\rangle|$ and then follows from Assumptions I and II and Lemma3.4.3.

Similarly,

$$\left\|\mathbf{U}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star)\mathbf{B}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^\star\right\| \le \left[2\kappa(\log n)\frac{\sqrt{r}}{n} + \frac{\tau(\log n)}{3\sqrt{n}}\right]\left(1 + c\sqrt{\frac{r}{n}}\right)^2.$$

The similar arguments also apply to bounding $|\mathbf{x}^\top(\mathbf{U}\,\mathrm{diag}(\mathbf{C}^\top\mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^\star)\mathbf{y}|$. Note that

$$\mathbf{x}^\top\left(\mathbf{U}^\star\,\mathrm{diag}(\mathbf{C}^\top\mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{u}_1^\star\mathbf{v}_1^{\star\top}\right)\mathbf{y} = \mathbf{x}^\top(\mathbf{U}\,\mathrm{diag}((\mathbf{C}-\mathbf{W}/3)^\top\mathbf{w}_1^\star)\mathbf{V}^\top)\mathbf{y} + \frac{1}{3}\mathbf{x}^\top(\mathbf{U}\,\mathrm{diag}(\mathbf{W}^\top\mathbf{w}_1^\star - \mathbf{e}_1)\mathbf{V}^\top)\mathbf{y}$$

and the first term can be rewritten as

$$\begin{aligned}\mathbf{x}^\top(\mathbf{U}\,\mathrm{diag}((\mathbf{C}-\mathbf{W}/3)^\top\mathbf{w}_1^\star)\mathbf{V}^\top)\mathbf{y} &= \sum_{i=1}^r \mathbf{x}^\top\left((\mathbf{c}_i - \mathbf{w}_i/3)^\top\mathbf{w}_1^\star\mathbf{u}_i\mathbf{v}_i^\top\right)\mathbf{y} \\ &= \sum_{i=1}^r(\mathbf{x}^\top\mathbf{u}_i)(\mathbf{v}_i^\top\mathbf{y})(\mathbf{c}_i - \mathbf{w}_i/3)^\top\mathbf{w}_1^\star) \\ &= \mathbf{x}^\top\sum_{i=1}^r\left(\mathbf{u}_i(\mathbf{v}_i^\top\mathbf{y})(\mathbf{c}_i - \mathbf{w}_i/3)^\top\right)\mathbf{w}_1^\star \\ &= \mathbf{x}^\top\left(\mathbf{U}\,\mathrm{diag}(\mathbf{V}^\top\mathbf{y})(\mathbf{C}-\mathbf{W}/3)^\top\right)\mathbf{w}_1^\star,\end{aligned}$$

and so

$$\left|\mathbf{x}^\top \left(\mathbf{U}^\star \operatorname{diag}(\mathbf{C}^\top \mathbf{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{u}_1^\star \mathbf{v}_1^{\star\top}\right)\mathbf{y}\right| \leq \|\mathbf{U}\|\|\operatorname{diag}(\mathbf{V}^\top \mathbf{y})\|\|\mathbf{C} - \mathbf{W}/3\| + \frac{1}{3}\|\mathbf{U}\|\|\operatorname{diag}(\mathbf{W}^\top \mathbf{w}_1^\star - \mathbf{e}_1)\|\|\mathbf{V}^\top\|.$$

Finally, we obtain

$$\begin{aligned}
|q(\mathbf{x}, \mathbf{y}, \mathbf{w}_1^\star)| \leq & \left[6\kappa(\log n)\frac{\sqrt{r}}{n} + \frac{\tau(\log n)}{\sqrt{n}}\right]\left(1 + c\sqrt{\frac{r}{n}}\right)^2 \\
= & O\left(\frac{\kappa(\log n)\sqrt{r}}{n}, \frac{\tau(\log n)}{\sqrt{n}}, \frac{\kappa(\log n) r^{1.5}}{n^2}, \frac{\tau(\log n) r}{n^{1.5}}\right) \\
= & O\left(\frac{\kappa(\log n)}{n^{3/8+\frac{3}{4}r_c}}, \frac{\tau(\log n)}{n^{5/8-\frac{3}{4}r_c}}, \frac{\kappa(\log n)}{n^{1/8+\frac{9}{4}r_c}}, \frac{\tau(\log n)}{n^{\frac{1}{4}+1.5r_c}}\right) \\
= & O(\kappa(\log n) n^{-3r_c}, \tau(\log n) n^{-3r_c}) = o(n^{-2r_c})
\end{aligned}$$

with the notation $O(f(n), g(n)) := \max\{O(f(n)), O(g(n))\}$. The the last second line holds if $r \leq n^{1.25-1.5r_c}$ and the last line follows from the assumption $r_c \in (0, 1/6)$.

The same bound holds for $|q(\mathbf{x}, \mathbf{v}_1^\star, \mathbf{z})|$ and $|q(\mathbf{u}_1^\star, \mathbf{y}, \mathbf{z})|$.

The coefficient of the last term of (3.37) is $q(\mathbf{x}, \mathbf{y}, \mathbf{z})$ and its absolute value is bounded by the tensor spectral norm of $\boldsymbol{\mathcal{Q}}$, and should be close to constant as $\boldsymbol{\mathcal{Q}}$ is close to $\sum_{p=1}^r \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star$, the spectral norm of which is $1 + O(n^{-r_c})$ by the following lemma.

Lemma B.5.2. *Under Assumptions I and II, and if $r \leq n^{1.25-1.5r_c}$ with $r_c \in (0, 1/6)$,*

$$\left\|\sum_{p=1}^r \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star\right\| \leq 1 + \frac{5}{4}\tau(\log n) n^{-r_c}.$$

Proof of LemmaB.5.2.

$$\begin{aligned}
\left\|\sum_{p=1}^r \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star\right\| = & \sup_{(\mathbf{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \langle \mathbf{U}^\top \mathbf{a}, (\mathbf{V}^\top \mathbf{b}) \odot (\mathbf{W}^\top \mathbf{c})\rangle \\
\leq & \sup_{(\mathbf{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \|\mathbf{U}^\top \mathbf{a}\|_3 \|(\mathbf{V}^\top \mathbf{b}) \odot (\mathbf{W}^\top \mathbf{c})\|_{3/2} \\
\leq & \sup_{(\mathbf{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \|\mathbf{U}^\top \mathbf{a}\|_3 \|\mathbf{V}^\top \mathbf{b}\|_3 \|\mathbf{W}^\top \mathbf{v}\|_3 \\
\leq & \|\mathbf{U}^\top\|_{2\to 3}\|\mathbf{V}^\top\|_{2\to 3}\|\mathbf{W}^\top\|_{2\to 3} \\
\leq & \left(1 + \frac{1}{3}\tau(\log n) n^{-r_c}\right)^3 \\
= & 1 + \tau(\log n) n^{-r_c} + \frac{1}{3}\tau(\log n)^2 n^{-r_c} + \frac{1}{9}\tau(\log n)^3 n^{-3r_c} \\
\leq & 1 + \frac{5}{4}\tau(\log n) n^{-r_c},
\end{aligned}$$

where the first inequality follows from *Hölder's* inequality and the second inequality follows from *Cauchy's* inequality. The fourth inequality follows from LemmaB.4.2 when $r \leq n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$. The last inequality holds

since $\frac{1}{3}\tau(\log n)^2 n^{-r_c} + \frac{1}{9}\tau(\log n)^3 n^{-3r_c} \ll \frac{1}{4}n^{-r_c}$. $\square$

It remains to bound the difference between $\boldsymbol{\mathcal{Q}}$ and $\sum_{p=1}^r \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star$:

$$\begin{aligned}\left\|\mathbf{Q} - \sum_{p=1}^r \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star\right\| \le & \underbrace{\left\|\sum_{p=1}^r (\boldsymbol{\alpha}_p^\star - \frac{1}{3}\mathbf{u}_p^\star) \otimes \mathbf{v}_p^\star \otimes \mathbf{w}_p^\star\right\|}_{\Pi_1} + \underbrace{\left\|\sum_{p=1}^r \mathbf{u}_p^\star \otimes (\boldsymbol{\beta}_p^\star - \frac{1}{3}\mathbf{v}_p^\star) \otimes \mathbf{w}_p^\star\right\|}_{\Pi_2} \\ & + \underbrace{\left\|\sum_{p=1}^r \mathbf{u}_p^\star \otimes \mathbf{v}_p^\star \otimes (\boldsymbol{\gamma}_p^\star - \frac{1}{3}\mathbf{w}_p^\star)\right\|}_{\Pi_3}.\end{aligned}$$

First we bound Π_1:

$$\begin{aligned}\Pi_1 = & \sup_{(\mathbf{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \langle (\mathbf{A} - \frac{1}{3}\mathbf{U})^\top \mathbf{a}, (\mathbf{V}^\top \mathbf{b}) \odot (\mathbf{W}^\top \mathbf{c})\rangle \\ \le & \sup_{(\mathbf{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \|(\mathbf{A} - \frac{1}{3}\mathbf{U})^\top \mathbf{x}\|_2 \|(\mathbf{V}^\top \mathbf{b}) \odot (\mathbf{W}^\top \mathbf{c})\|_2 \\ \le & \sup_{(\mathbf{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \|(\mathbf{A} - \frac{1}{3}\mathbf{U})^\top \mathbf{x}\|_2 \|(\mathbf{V}^\top \mathbf{b})\|_4 \|(\mathbf{W}^\top \mathbf{c})\|_4 \\ \le & \|\mathbf{A} - \frac{1}{3}\mathbf{U}\| \|\mathbf{V}^\top\|_{2\to 4} \|\mathbf{W}^\top\|_{2\to 4} \\ \le & 2\kappa(\log n)\left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}}\right)(1 + o(1)) \le 8\kappa(\log n)\max\left\{\frac{\sqrt{r}}{n}, c\frac{r}{n^{1.5}}\right\} \le 8\kappa(\log n) n^{-3r_c} = o(n^{-2r_c})\end{aligned}$$

where the first and second inequalities follows from *Cauchy's* inequality and the fourth inequality follows from Lemma3.4.3 and LemmaB.4.2 when $r \ll n^{1.25}$. The last inequality follows by plugging $r \le n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$.

The same bound also holds for Π_2 and Π_3.

Combine ALL. If $r \le n^{1.25-1.5r_c}$ with $r_c \in (0, 1/6)$, we have

$$\begin{aligned}&|q(\mathbf{u}_1^\star, \mathbf{y}, \mathbf{z})| = o(n^{-2r_c}), \\ &|q(\mathbf{x}, \mathbf{v}_1^\star, \mathbf{z})| = o(n^{-2r_c}), \\ &|q(\mathbf{x}, \mathbf{y}, \mathbf{w}_1^\star)| = o(n^{-2r_c}), \\ &|q(\mathbf{x}, \mathbf{y}, \mathbf{z})| \le 1 + \frac{5}{4}\tau(\log n) n^{-r_c} + o(n^{-2r_c}),\end{aligned} \tag{B.20}$$

which together with (3.37) gives

$$\begin{aligned}&|q(\mathbf{u}(\theta_1), \mathbf{v}(\theta_2), \mathbf{w}(\theta_3))| \\ \le & |\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)| + |\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)| + \frac{5}{4}\tau(\log n) n^{-r_c} + o(n^{-2r_c}) \\ \le & |\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)| + |\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)| + \frac{4}{3}\tau(\log n) n^{-r_c}\end{aligned}$$

where the last inequality follows from $o(n^{-2r_c}) \ll \frac{1}{12}\tau(\log n)n^{-r_c}$.

□

B.6 Proof of Lemma 3.4.6

Lemma B.6.1 (Lemma 3.4.6)**.** *Under Assumptions I, II, III, if $r \ll n^{1.25}$, then for any $\xi_i \in \left(-\frac{\sqrt{2}-1}{3}, \frac{\sqrt{2}-1}{3}\right)$, we have*

$$F(\theta_1+\xi_1, \theta_2+\xi_2, \theta_3+\xi_3) \leq 1 \tag{3.42}$$

for $(\theta_1,\theta_2,\theta_3) \in \{(0,0,0),(0,\pi,\pi),(\pi,0,\pi),(\pi,\pi,0)\}$ and

$$F(\theta_1+\xi_1, \theta_2+\xi_2, \theta_3+\xi_3) < 0 \tag{3.43}$$

for $(\theta_1,\theta_2,\theta_3) \in \{(\pi,\pi,\pi),(\pi,0,0),(0,\pi,0),(0,0,\pi)\}$. Here, equality in (3.42) *holds only if $\xi_1=\xi_2=\xi_3=0$.*

Proof. Recall that

$$\begin{aligned} F(\theta_1,\theta_2,\theta_3) = \cos(\theta_1)\cos(\theta_2)\cos(\theta_3) &+ q(\mathbf{u}_1^\star,\mathbf{y},\mathbf{z})\cos(\theta_1)\sin(\theta_2)\sin(\theta_3) \\ &+ q(\mathbf{x},\mathbf{v}_1^\star,\mathbf{z})\sin(\theta_1)\cos(\theta_2)\sin(\theta_3) \\ &+ q(\mathbf{x},\mathbf{y},\mathbf{w}_1^\star)\sin(\theta_1)\sin(\theta_2)\cos(\theta_3) \\ &+ q(\mathbf{x},\mathbf{y},\mathbf{z})\sin(\theta_1)\sin(\theta_2)\sin(\theta_3). \end{aligned} \tag{B.21}$$

The points of special interest are the eight vertices of the cube $[0,\pi]\times[0,\pi]\times[0,\pi]$, *i.e.,*

$$\{(\theta_1,\theta_2,\theta_3) : \theta_i \in \{0,\pi\}, i=1,2,3\}$$

which we classify into two sets:

- The first set of vertices involve an even number of π: $(0,0,0),(0,\pi,\pi),(\pi,0,\pi),(\pi,\pi,0)$;
- The second set of vertices involve an odd number of π: $(\pi,0,0),(0,\pi,0),(0,0,\pi),(\pi,\pi,\pi)$.

Controlling the first vertex set. For the first set of points, we only show that

$$F(\theta_1+\xi_1,\theta_2+\xi_2,\theta_3+\xi_3) \leq 1, \quad \forall \xi_i \in \left(-\frac{\sqrt{2}-1}{3}, \frac{\sqrt{2}-1}{3}\right) \bigcup \left(\frac{\pi}{2}-\frac{\sqrt{2}-1}{3}, \frac{\pi}{2}+\frac{\sqrt{2}-1}{3}\right)$$

holds for $(\theta_1,\theta_2,\theta_3)=(0,0,0)$. The same arguments apply to the other cases $(\pi,0,\pi),(0,\pi,\pi),(\pi,\pi,0)$ since (B.21) implies

$$F(\xi_1,\xi_2,\xi_3) = F(\xi_1,\pi+\xi_2,\pi+\xi_3) = F(\pi+\xi_1,\xi_2,\pi+\xi_3) = F(\pi+\xi_1,\pi+\xi_2,\xi_3)$$

for all $\xi_1,\xi_2\,\xi_3 \in \mathbb{R}$.

Let us apply the first-order Taylor expansion to $F(\theta_1, \theta_2, \theta_3)$ over some smaller cube $[-\theta_0, \theta_0] \times [-\theta_0, \theta_0] \times [-\theta_0, \theta_0]$ with $\theta_0 \in (0, \pi/2)$ to be determined later,

$$\begin{aligned} F(\theta_1, \theta_2, \theta_3) =& F(0,0,0) + \boldsymbol{\theta}^\top \nabla F(\xi_1, \xi_2, \xi_3) \\ \geq& 1 - \|\boldsymbol{\theta}\|_1 \sup_{|\xi_1|, |\xi_2|, |\xi_3| \leq \theta_0} \|\nabla F(\xi_1, \xi_2, \xi_3)\|_\infty, \end{aligned}$$

where $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \end{bmatrix}^\top$. Since

$$\begin{aligned} \frac{\partial}{\partial \theta_1} F(\xi_1, \xi_2, \xi_3) = &- \sin(\xi_1)\cos(\xi_2)\cos(\xi_3) - q(\mathbf{u}_1^\star, \mathbf{y}, \mathbf{z})\sin(\xi_1)\sin(\xi_2)\sin(\xi_3) \\ &+ q(\mathbf{x}, \mathbf{v}_1^\star, \mathbf{z})\cos(\xi_1)\cos(\xi_2)\sin(\xi_3) \\ &+ q(\mathbf{x}, \mathbf{y}, \mathbf{w}_1^\star)\cos(\xi_1)\sin(\xi_2)\cos(\xi_3) \\ &+ q(\mathbf{x}, \mathbf{y}, \mathbf{z})\cos(\xi_1)\sin(\xi_2)\sin(\xi_3), \end{aligned}$$

we have

$$\begin{aligned} \left| \frac{\partial}{\partial \theta_1} F(\xi_1, \xi_2, \xi_3) \right| \leq& |\sin(\theta_0)| + o(1)(|\sin(\theta_0)|^3 + 2|\sin(\theta_0)|) + (1 + o(1))|\sin(\theta_0)|^2 \\ \leq& |\sin(\theta_0)| + |\sin(\theta_0)|^2 + o(1) \\ \leq& 3|\sin(\theta_0)| \end{aligned} \tag{B.22}$$

where the first inequality follows from (B.20), and so

$$|q(\mathbf{u}_1^\star, \mathbf{y}, \mathbf{z})| = o(1), \quad |q(\mathbf{x}, \mathbf{v}_1^\star, \mathbf{z})| = o(1), \quad |q(\mathbf{x}, \mathbf{y}, \mathbf{w}_1^\star)| = o(1), \quad |q(\mathbf{x}, \mathbf{y}, \mathbf{z})| = 1 + o(1) \tag{B.23}$$

under Assumptions I-III and $r \ll n^{1.25}$ (by letting r_c in "$r \ll n^{1.25-r_c}$" approach to zero). The inequality (B.22) uses the facts that $|\sin(\theta_0)|^2 \leq |\sin(\theta_0)|$ and $o(1) \leq |\sin(\theta_0)|$ for sufficiently large n. The same bound holds for $\left|\frac{\partial}{\partial \theta_2} F(\xi_1, \xi_2, \xi_3)\right|$ and $\left|\frac{\partial}{\partial \theta_3} F(\xi_1, \xi_2, \xi_3)\right|$. We therefore have

$$F(\theta_1, \theta_2, \theta_3) \geq 1 - 3\|\boldsymbol{\theta}\|_1 |\sin(\theta_0)| \geq 1 - 9\theta_0^2. \tag{B.24}$$

Let us compute the second-order Taylor expansion of $F(\theta_1, \theta_2, \theta_3)$:

$$F(\theta_1, \theta_2, \theta_3) = F(0,0,0) + \boldsymbol{\theta}^\top \nabla F(0,0,0) + \frac{1}{2}\boldsymbol{\theta}^\top \nabla^2 F(\xi_1, \xi_2, \xi_3)\boldsymbol{\theta}$$

where $(\xi_1, \xi_2, \xi_3) \in [-\theta_0, \theta_0]^3$. As a consequence of the construction process of the dual polynomial, we have $F(0,0,0) = 1$ and $\nabla F(0,0,0) = 0$, implying

$$F(\theta_1, \theta_2, \theta_3) = 1 + \frac{1}{2}\boldsymbol{\theta}^\top \nabla^2 F(\xi_1, \xi_2, \xi_3)\boldsymbol{\theta}.$$

Therefore, as long as we can find θ_0 such that the Hessian matrix $\nabla^2 F$ is negative definite over the region $[-\theta_0, \theta_0]^3$, then $F(\theta_1, \theta_2, \theta_3) \leq 1$ for any $(\theta_1, \theta_2, \theta_3) \in [-\theta_0, \theta_0]^3$ with equality holds only if $(\theta_1, \theta_2, \theta_3) = (0, 0, 0)$.

We next estimate the Hessian matrix $\nabla^2 F(\xi_1, \xi_2, \xi_3)$. Direct computation gives

$$\nabla^2 F(\xi_1, \xi_2, \xi_3) = \begin{bmatrix} -F(\xi_1, \xi_2, \xi_3) & * & * \\ * & -F(\xi_1, \xi_2, \xi_3) & * \\ * & * & -F(\xi_1, \xi_2, \xi_3) \end{bmatrix}$$

whose off-diagonal elements are nonsymmetric partial derivatives of F, for example,

$$\begin{aligned} \frac{\partial^2}{\partial\theta_1\partial\theta_2} F(\theta_1, \theta_2, \theta_3) = \sin(\theta_1)\sin(\theta_2)\cos(\theta_3) &- q(\mathbf{u}_1^\star, \mathbf{y}, \mathbf{z})\sin(\theta_1)\cos(\theta_2)\sin(\theta_3) \\ &+ q(\mathbf{x}, \mathbf{y}, \mathbf{w}_1^\star)\cos(\theta_1)\cos(\theta_2)\cos(\theta_3) \\ &- q(\mathbf{x}, \mathbf{v}_1^\star, \mathbf{z})\cos(\theta_1)\sin(\theta_2)\sin(\theta_3) \\ &+ q(\mathbf{x}, \mathbf{y}, \mathbf{z})\cos(\theta_1)\cos(\theta_2)\sin(\theta_3), \end{aligned}$$

which implies by (B.23) that

$$\begin{aligned} \left| \frac{\partial^2}{\partial\theta_1\partial\theta_2} F(\theta_1, \theta_2, \theta_3) \right| \leq & |\sin(\theta_0)|^2 + o(1)(1 + 2|\sin(\theta_0)|^2) + (1 + o(1))|\sin(\theta_0)| \\ \leq & |\sin(\theta_0)| + |\sin(\theta_0)|^2 + o(1) \\ \leq & 3|\sin(\theta_0)|. \end{aligned}$$

The same bound holds for other mixed partial derivatives $\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j} F(\theta_1, \theta_2, \theta_3)\right|$ with $i, j = 1, 2, 3$ and $i \neq j$.

To make $\nabla^2 F(\xi_1, \xi_2, \xi_3)$ negative definite, by *Gershgorin's circle theorem* and the bound (B.24), we only need

$$-F(\xi_1, \xi_2, \xi_3) + 6|\sin(\theta_0)| \leq -1 + 9\theta_0^2 + 6\theta_0 < 0$$

which holds for $\theta_0 \in (\frac{-\sqrt{2}-1}{3}, \frac{\sqrt{2}-1}{3})$, including $(\frac{-\sqrt{2}+1}{3}, \frac{\sqrt{2}-1}{3})$. This completes the first part of the proof.

Controlling the second vertex set. Similarly as before, we first show

$$F(\pi + \xi_1, \pi + \xi_2, \pi + \xi_3) < 0, \ \forall |\xi_i| < \frac{\sqrt{2} - 1}{3}.$$

It follows from the intermediate result (B.24):

$$F(\xi_1, \xi_2, \xi_3) \geq 1 - 9\theta_0^2 > 0, \ \forall |\xi_i| \leq \theta_0$$

by recognizing that $F(\pi + \xi_1, \pi + \xi_2, \pi + \xi_3) = -F(\xi_1, \xi_2, \xi_3), \forall \xi_1, \xi_2, \xi_3$ and choosing $\theta_0 = (\sqrt{2} - 1)/3$. Finally, we claim the same conclusion applies to the remaining three cases since

$$F(\pi + \xi_1, \pi + \xi_2, \pi + \xi_3) = F(\pi + \xi_1, \xi_2, \xi_3) = F(\xi_1, \pi + \xi_2, \xi_3) = F(\xi_1, \xi_2, \pi + \xi_3)$$

for all $\xi_1, \xi_2, \xi_3 \in \mathbb{R}$.

□

B.7 Proof of Lemma 3.4.7

Lemma B.7.1 (Lemma 3.4.7)**.** *Under Assumptions I, II, III, if $r \leq n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$, then for sufficiently large n, we have $|F(\theta_1, \theta_2, \theta_3)| < 1$ in $\mathbb{N}_b(\delta_b)$ for $\delta_b = \sqrt{\frac{80\tau(\log n)}{3}} n^{-0.5r_c}$.*

Proof. First, solve for θ such that

$$|\cos(\theta)^3| + |\sin(\theta)|^3 < 1 - 4\tau(\log n)n^{-r_c}. \tag{B.25}$$

To this end, we define $f(\theta) := |\cos(\theta)^3| + |\sin(\theta)|^3$ for $\theta \in [0, \pi]$. It can be verified directly that f is symmetric around $\frac{\pi}{2}$ on $[0, \pi]$, symmetric around $\frac{\pi}{4}$ on $[0, \frac{\pi}{2}]$, and strictly decreasing on $[0, \frac{\pi}{4}]$. Since $1 - 4\tau(\log n)n^{-r_c} \in (0, 1)$, there exists a unique $\varpi \in (0, \frac{\pi}{4})$ such that $f(\varpi) = 1 - 4\tau(\log n)n^{-r_c} \in (0, 1)$. Thus the inequality (B.25) holds on $(\varpi, \frac{\pi}{2} - \varpi) \cup (\frac{\pi}{2} + \varpi, \pi - \varpi)$.

To have an approximation of ϖ, we need the following lemma.

Lemma B.7.2. *Let f and g be any two real functions with g being strictly decreasing in some interval (α, β) and satisfying $g(x) \geq f(x), \forall x \in (\alpha, \beta)$. Suppose both equations $f(x) = b$ and $g(x) = b$ admit one root in $[\alpha, \beta]$, denoted by x_f and x_g respectively. Then $x_g \geq x_f$.*

Proof of Lemma B.7.2. Since $g(x) > g(x_f) \geq f(x_f) = b$ for any $x \in [\alpha, x_f)$, $g(x_g) = b$ could only happen within $[x_f, \beta]$. □

We now recognize that

$$f(\theta) \leq 1 - \frac{3}{20}\theta^2, \text{ for } \theta \in [0, \pi/4] \tag{B.26}$$

and $g(\theta) := 1 - \frac{3}{20}\theta^2$ is strictly deceasing $[0, \pi/4]$. Clearly,

$$\delta_b := \sqrt{\frac{80\tau(\log n)}{3}} n^{-0.5r_c}$$

is the root of $g(\theta) = 1 - 4\tau(\log n)n^{-r_c}$ over the interval $[0, \frac{\pi}{4}]$. By Lemma B.7.2, $\delta_b \geq \varpi$. Therefore, (B.25) holds on $(\delta_b, \frac{\pi}{2} - \delta_b) \cup (\frac{\pi}{2} + \delta_b, \pi - \delta_b)$. By (3.47), we obtain

$$F(\theta_1, \theta_2, \theta_3) < 1 \text{ for } (\theta_1, \theta_2, \theta_3) \in \mathbb{N}_b(\delta_b).$$

□

B.7.1 Proof of Eq. (B.26)

Showing (B.26) is equivalent to showing

$$\sin^3(x) + \cos^3(x) \leq 1 - \frac{3}{20}x^2, \ \forall x \in [0, \pi/4] \tag{B.27}$$

since $\sin(x), \cos(x) > 0$ for $x \in [0, \pi/4]$. Before moving on, we need the following lemma to prove (B.27).

Lemma B.7.3. *The following inequality*

$$\frac{(3^{2n-1}-3)}{4\cdot(2n-1)!}x^{2n-1} + \frac{(3^{2n}+3)}{4\cdot(2n)!}x^{2n} - \frac{(3^{2n+1}-3)}{4\cdot(2n+1)!}x^{2n+1} - \frac{(3^{2n+2}+3)}{4\cdot(2n+2)!}x^{2n+2} \geq 0 \tag{B.28}$$

holds for all $x \in [0, \pi/4]$ and $n \geq 2$,

Proof of Lemma B.7.3. Let p equal the expression on the left side of Equation (B.28). A simplification on p yields

$$p(x) = q_1(x)\frac{x^{2n-1}}{4(2n-1)!} + q_2(x)\frac{x^{2n+2}}{4(2n)!},$$

where $q_1(x) = (3^{2n-1}-3) - \dfrac{3^{2n+1}-3}{2n(2n+1)}x^2$ and $q_2(x) = (3^{2n}+3) - \dfrac{3^{2n+2}+3}{(2n+1)(2n+2)}x^2$.

As functions of x, q_1 and q_2 have roots at

$$\pm\sqrt{\frac{2n(2n+1)(3^{2n-1}-3)}{3^{2n+1}-3}} \quad \text{and} \quad \pm\sqrt{\frac{(2n+1)(2n+2)(3^{2n}+3)}{3^{2n+2}+3}},$$

respectively, provided $n \geq 1$. Since $10(3^{2n-1}-3) \geq 3^{2n+1}-3$ and $9(3^{2n}+3) > (3^{2n+2}+3)$ for all $n \geq 2$, it follows that the positive root of q_1 satisfies

$$\sqrt{\frac{2n(2n+1)(3^{2n-1}-3)}{3^{2n+1}-3}} \geq \sqrt{\frac{2n(2n+1)}{10}} > \sqrt{2} > \frac{\pi}{4}, \text{ for } n \geq 2,$$

and the positive root of q_2 satisfies

$$\sqrt{\frac{(2n+1)(2n+2)(3^{2n}+3)}{3^{2n+2}+3}} > \sqrt{\frac{(2n+1)(2n+2)}{9}} > \sqrt{\frac{10}{3}} > \frac{\pi}{4}, \text{ for } n \geq 2.$$

Therefore both q_1 and q_2 are positive on $[0, \pi/4]$ for all $n \geq 2$, and Equation (B.28) holds. □

B.7.1.1 The Proof

Lemma B.7.4. *The following statement*

$$\sin^3(x) + \cos^3(x) \leq 1 - \frac{3}{20}x^2$$

holds for all $x \in [0, \frac{\pi}{4}]$.

Proof. Recall that $\sin^3(x) = \frac{1}{4}(3\sin(x) - \sin(3x))$ and $\cos^3(x) = \frac{1}{4}(3\cos(x) + \cos(3x))$, and therefore

$$\sin^3(x) = x^3 + \sum_{n=5}^{\infty}(-1)^n \frac{3^{2n-1}-3}{4(2n-1)!}x^{2n-1},$$

and

$$\cos^3(x) = 1 - \frac{3}{2}x^2 + \frac{7}{8}x^4 + \sum_{n=3}^{\infty}(-1)^n \frac{3^{2n}+3}{4(2n)!}x^{2n}.$$

Thus

$$\sin^3(x) + \cos^3(x) \leq 1 - \frac{3}{2}x^2 + x^3 + \frac{7}{8}x^4,$$

for all $x \in [0, \pi/4]$ since by Lemma B.7.3

$$\begin{aligned}
&\sum_{n=3}^{\infty}(-1)^n \frac{3^{2n-1}-3}{4(2n-1)!}x^{2n-1} + \sum_{n=3}^{\infty}(-1)^n \frac{3^{2n}+3}{4(2n)!}x^{2n} \\
&= -\sum_{n=3,\ n \text{ odd}}^{\infty}\left(\frac{3^{2n-1}-3}{4(2n-1)!}x^{2n-1} + \frac{3^{2n}+3}{4(2n)!}x^{2n} - \frac{3^{2n+1}-3}{4(2n+1)!}x^{2n+1} - \frac{3^{2n+2}}{4(2n+2)!}x^{2n+2}\right) \leq 0.
\end{aligned}$$

Finally, note that

$$1 - \frac{3}{2}x^2 + x^3 + \frac{7}{8}x^4 = 1 - \frac{3}{20}x^2 + x^2 h(x),$$

with

$$h(x) = -\frac{27}{20} + x + \frac{7}{8}x^2$$

being negative in $[0, \pi/4]$. So the proof is complete. □

APPENDIX C

APPENDICES FOR CHAPTER 4

C.1 Proof of Proposition 4.3.1

Proposition C.1.1 (Proposition 4.3.1)**.** *Under the same setting as in Theorem 4.3.1, for any initial point* $\mathbf{U}_0$*,* $g(\mathbf{U})$ *on* $\mathrm{Lev}_f(\mathbf{U}_0)$ *defined in* (4.10) *has a Lipschitz continuous gradient with the Lipschitz constant*

$$L_c = \sqrt{2\beta\sqrt{\frac{2}{\alpha}(f(\mathbf{U}_0\mathbf{U}_0^\top) - f(\mathbf{X}^\star))} + 2\|\nabla f(\mathbf{X}^\star)\|_F + 4\beta\left(\|\mathbf{U}^\star\|_F + \frac{\sqrt{\frac{2}{\alpha}(f(\mathbf{U}_0\mathbf{U}_0^\top) - f(\mathbf{X}^\star))}}{2(\sqrt{2}-1)\rho(\mathbf{U}^\star)}\right)^2},$$

where $\rho(\cdot)$ *denotes the smallest nonzero singular value of its argument.*

Proof. To that end, we first show that for any $\mathbf{U} \in \mathrm{Lev}_f(\mathbf{U}_0)$, $\|\mathbf{U}\|_F$ is upper-bounded. Let $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$ and consider the following second-order Taylor expansion of $f(\mathbf{X})$

$$\begin{aligned} f(\mathbf{X}) &= f(\mathbf{X}^\star) + \langle \nabla f(\mathbf{X}^\star), \mathbf{X} - \mathbf{X}^\star\rangle + \frac{1}{2}\int_0^1 [\nabla^2 f(t\mathbf{X}^\star + (1-t)\mathbf{X})](\mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star)dt \\ &\geq f(\mathbf{X}^\star) + \frac{1}{2}\int_0^1 [\nabla^2 f(t\mathbf{X}^\star + (1-t)\mathbf{X})](\mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star)dt \\ &\geq f(\mathbf{X}^\star) + \frac{\alpha}{2}\|\mathbf{X} - \mathbf{X}^\star\|_F^2, \end{aligned}$$

which implies that

$$\|\mathbf{U}\mathbf{U}^\top - \mathbf{X}^\star\|_F^2 \leq \frac{2}{\alpha}(f(\mathbf{U}\mathbf{U}^\top) - f(\mathbf{X}^\star)) \leq \frac{2}{\alpha}(f(\mathbf{U}_0\mathbf{U}_0^\top) - f(\mathbf{X}^\star)) \tag{C.1}$$

with the second inequality following from the assumption $U \in \mathrm{Lev}_f(\mathbf{U}_0)$. Thus, we have

$$\|\mathbf{U}\|_F \leq \|\mathbf{U}^\star\|_F + \mathrm{dist}(\mathbf{U}, \mathbf{U}^\star) \leq \|\mathbf{U}^\star\|_F + \frac{\|\mathbf{U}\mathbf{U}^\top - \mathbf{X}^\star\|_F}{2(\sqrt{2}-1)\rho(\mathbf{U}^\star)} \leq \|\mathbf{U}^\star\|_F + \frac{\sqrt{\frac{2}{\alpha}(f(\mathbf{U}_0\mathbf{U}_0^\top) - f(\mathbf{X}^\star))}}{2(\sqrt{2}-1)\rho(\mathbf{U}^\star)}. \tag{C.2}$$

Now we are ready to show the Lipschitz gradient for g at $\mathrm{Lev}_f(\mathbf{U}_0)$:

$$
\begin{aligned}
\|\nabla^2 g(\mathbf{U})\|^2 &= \max_{\|\mathbf{D}\|_F=1} \left|[\nabla^2 g(\mathbf{U})](\mathbf{D},\mathbf{D})\right| \\
&= \max_{\|\mathbf{D}\|_F=1} \left|2\langle \nabla f(\mathbf{U}\mathbf{U}^\top), \mathbf{D}\mathbf{D}^\top\rangle + [\nabla^2 f(\mathbf{U}\mathbf{U}^\top)](\mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top, \mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top)\right| \\
&\le 2 \max_{\|\mathbf{D}\|_F=1} \left|\langle \nabla f(\mathbf{U}\mathbf{U}^\top), \mathbf{D}\mathbf{D}^\top\rangle\right| + \max_{\|\mathbf{D}\|_F=1} \left|[\nabla^2 f(\mathbf{U}\mathbf{U}^\top)](\mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top, \mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top)\right| \\
&\le 2 \max_{\|\mathbf{D}\|_F=1} \left|\langle \nabla f(\mathbf{U}\mathbf{U}^\top) - \nabla f(\mathbf{X}^\star), \mathbf{D}\mathbf{D}^\top\rangle\right| + 2\|\nabla f(\mathbf{X}^\star)\|_F + \beta\|\mathbf{D}\mathbf{U}^\top + \mathbf{U}\mathbf{D}^\top\|_F^2 \\
&\le 2\beta\|\mathbf{U}\mathbf{U}^\top - \mathbf{X}^\star\|_F + 2\|\nabla f(\mathbf{X}^\star)\|_F + 4\beta\|\mathbf{U}\|_F^2 \\
&\le 2\beta\sqrt{\frac{2}{\alpha}(f(\mathbf{U}_0\mathbf{U}_0^\top) - f(\mathbf{X}^\star))} + 2\|\nabla f(\mathbf{X}^\star)\|_F + 4\beta\left(\|\mathbf{U}^\star\|_F + \frac{\sqrt{\frac{2}{\alpha}(f(\mathbf{U}_0\mathbf{U}_0^\top) - f(\mathbf{X}^\star))}}{2(\sqrt{2}-1)\rho(\mathbf{U}^\star)}\right)^2 \\
&:= L_c^2.
\end{aligned}
$$

Here, the last second line follows from (C.1) and (C.2). This concludes the proof of Proposition 4.3.1. □

C.2 Proof of Lemma 4.3.2

Lemma C.2.1 (Lemma 4.3.2)**.** *Assume that* $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{n\times r}$*. Then*

$$
\|\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{U}_2\mathbf{U}_2^\top\|_F \ge \min\{\rho(\mathbf{U}_1), \rho(\mathbf{U}_2)\}\,\mathrm{dist}(\mathbf{U}_1, \mathbf{U}_2).
$$

Proof. Let $\mathbf{X}_1 = \mathbf{U}_1\mathbf{U}_1^\top$, $\mathbf{X}_2 = \mathbf{U}_2\mathbf{U}_2^\top$ and their full eigenvalue decompositions be

$$
\mathbf{X}_1 = \sum_{j=1}^n \lambda_j \mathbf{p}_j\mathbf{p}_j^\top, \qquad \mathbf{X}_2 = \sum_{j=1}^n \eta_j \mathbf{q}_j\mathbf{q}_j^\top
$$

where $\{\lambda_j\}$ and $\{\eta_j\}$ are the eigenvalues in decreasing order. Since $\mathrm{rank}(\mathbf{U}_1) = r_1$ and $\mathrm{rank}(\mathbf{U}_2) = r_2$, we have $\lambda_j = 0$ for $j > r_1$ and $\eta_j = 0$ for $j > r_2$. We compute $\|\mathbf{X}_1 - \mathbf{X}_2\|_F^2$ as follows

$$
\begin{aligned}
\|\mathbf{X}_1 - \mathbf{X}_2\|_F^2 &= \|\mathbf{X}_1\|_F^2 + \|\mathbf{X}_2\|_F^2 - 2\langle \mathbf{X}_1, \mathbf{X}_2\rangle \\
&= \sum_{i=1}^n \lambda_i^2 + \sum_{j=1}^n \eta_j^2 - \sum_{i=1}^n\sum_{j=1}^n 2\lambda_i\eta_j\langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 \\
&\overset{①}{=} \sum_{i=1}^n \lambda_i^2 \sum_{j=1}^n \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 + \sum_{j=1}^n \eta_j^2 \sum_{i=1}^n \langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 - \sum_{i=1}^n\sum_{j=1}^n 2\lambda_i\eta_j\langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 \\
&\overset{②}{=} \sum_{i=1}^n\sum_{j=1}^n (\lambda_i - \eta_j)^2\langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 \\
&= \sum_{i=1}^n\sum_{j=1}^n \left(\sqrt{\lambda_i} - \sqrt{\eta_j}\right)^2\left(\sqrt{\lambda_i} + \sqrt{\eta_j}\right)^2\langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 \\
&\overset{③}{\ge} \min\left\{\sqrt{\lambda_{r_1}}, \sqrt{\eta_{r_2}}\right\}^2 \sum_{i=1}^n\sum_{j=1}^n \left(\sqrt{\lambda_i} - \sqrt{\eta_j}\right)^2\langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 \\
&\overset{④}{=} \min\{\lambda_{r_1}, \eta_{r_2}\}\left\|\sqrt{\mathbf{X}_1} - \sqrt{\mathbf{X}_2}\right\|_F^2,
\end{aligned}
$$

where ① uses the fact $\sum_{j=1}^{n}\langle \mathbf{p}_i, \mathbf{q}_j\rangle^2 = \|\mathbf{p}_i\|_2^2 = 1$ with $\{\mathbf{q}_j\}$ being an orthonormal basis and similarly $\sum_{i=1}^{n}\langle \mathbf{p}_i, \mathbf{q}_j\rangle^2$ $= \|\mathbf{q}_j\|_2^2 = 1$. ② is by firstly an exchange of the summations, secondly the fact that $\lambda_j = 0$ for $j > r_1$ and $\eta_j = 0$ for $j > r_2$, and thirdly completing squares. ③ is because $\{\lambda_j\}$ and $\{\eta_j\}$ are sorted in decreasing order. ④ follows from ② and that $\{\sqrt{\lambda_j}\}$ and $\{\sqrt{\eta_j}\}$ are eigenvalues of $\sqrt{\mathbf{X}_1}$ and $\sqrt{\mathbf{X}_2}$, the matrix square root of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

Finally, we can conclude the proof as long as we can show the following inequality:

$$\left\|\sqrt{\mathbf{X}_1} - \sqrt{\mathbf{X}_2}\right\|_F^2 \geq \min_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top = \mathbf{I}_r} \|\mathbf{U}_1 - \mathbf{U}_2\mathbf{R}\|_F^2. \tag{C.3}$$

By expanding $\|\cdot\|_F^2$ in (C.3) and noting that $\langle\sqrt{\mathbf{X}_1}, \sqrt{\mathbf{X}_1}\rangle = \mathrm{tr}(\mathbf{X}_1) = \mathrm{tr}(\mathbf{U}_1\mathbf{U}_1^\top)$ and $\langle\sqrt{\mathbf{X}_2}, \sqrt{\mathbf{X}_2}\rangle = \mathrm{tr}(\mathbf{X}_2) =$ $\mathrm{tr}(\mathbf{U}_2\mathbf{U}_2^\top)$, (C.3) reduces to

$$\langle\sqrt{\mathbf{X}}_1, \sqrt{\mathbf{X}_2}\rangle \leq \max_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top = \mathbf{I}_r} \langle \mathbf{U}_1, \mathbf{U}_2\mathbf{R}\rangle. \tag{C.4}$$

To show (C.4), we write the SVDs of $\mathbf{U}_1, \mathbf{U}_2$ respectively as $\mathbf{U}_1 = \mathbf{P}_1\mathbf{\Sigma}_1\mathbf{Q}_1^\top$ and $\mathbf{U}_2 = \mathbf{P}_2\mathbf{\Sigma}_2\mathbf{Q}_2^\top$ with $\mathbf{P}_1, \mathbf{P}_2 \in$ $\mathbb{R}^{n\times r}$, $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \mathbb{R}^{r\times r}$ and $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathbb{R}^{r\times r}$. Then we have $\sqrt{\mathbf{X}_1} = \mathbf{P}_1\mathbf{\Sigma}_1\mathbf{P}_1^\top$, $\sqrt{\mathbf{X}_2} = \mathbf{P}_2\mathbf{\Sigma}_2\mathbf{P}_2^\top$.

On one hand,

$$\begin{aligned}
\text{RHS of (C. 4)} &= \max_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top = \mathbf{I}_r} \left\langle \mathbf{P}_1\mathbf{\Sigma}_1\mathbf{Q}_1^\top, \mathbf{P}_2\mathbf{\Sigma}_2\mathbf{Q}_2^\top\mathbf{R}\right\rangle \\
&= \max_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top = \mathbf{I}_r} \left\langle \mathbf{P}_1\mathbf{\Sigma}_1, \mathbf{P}_2\mathbf{\Sigma}_2, \mathbf{Q}_2^\top\mathbf{R}\mathbf{Q}_1\right\rangle \\
&= \max_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top = \mathbf{I}_r} \langle \mathbf{P}_1\mathbf{\Sigma}_1, \mathbf{P}_2\mathbf{\Sigma}_2\mathbf{R}\rangle && \text{By } \mathbf{R} \leftarrow \mathbf{Q}_2^\top\mathbf{R}\mathbf{Q}_1 \\
&= \|(\mathbf{P}_2\mathbf{\Sigma}_2)^\top\mathbf{P}_1\mathbf{\Sigma}_1\|_*. && \text{By Lemma 4.3.1}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\text{LHS of C. 4} &= \langle \mathbf{P}_1\mathbf{\Sigma}_1\mathbf{P}_1^\top, \mathbf{P}_2\mathbf{\Sigma}_2\mathbf{P}_2^\top\rangle \\
&= \langle (\mathbf{P}_2\mathbf{\Sigma}_2)^\top\mathbf{P}_1\mathbf{\Sigma}_1, \mathbf{P}_2^\top\mathbf{P}_1\rangle \\
&\leq \|(\mathbf{P}_2\mathbf{\Sigma}_2)^\top\mathbf{P}_1\mathbf{\Sigma}_1\|_*\|\mathbf{P}_2^\top\mathbf{P}_1\| && \text{By Hölder's Inequality} \\
&\leq \|(\mathbf{P}_2\mathbf{\Sigma}_2)^\top\mathbf{P}_1\mathbf{\Sigma}_1\|_*. && \text{Since } \|\mathbf{P}_2^\top\mathbf{P}_1\| \leq \|\mathbf{P}_2\|\|\mathbf{P}_1\| \leq 1
\end{aligned}$$

This proves (C.4) and hence completes the proof of Lemma 4.3.2.

□

C.3 Proof of Lemma 4.3.4

Lemma C.3.1 (Lemma 4.3.4)**.** *Let* $\mathbf{U}$ *and* $\mathbf{Z}$ *be any two matrices in* $\mathbb{R}^{n\times r}$ *such that* $\mathbf{U}^\top\mathbf{Z} = \mathbf{Z}^\top\mathbf{U}$ *is PSD. Assume that* $\mathbf{Q}$ *is an orthogonal matrix whose columns span* $\mathrm{Range}(\mathbf{U})$*. Then*

$$\left\|(\mathbf{U} - \mathbf{Z})\mathbf{U}^\top\right\|_F^2 \leq \frac{1}{8}\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top\right\|_F^2 + \left(3 + \frac{1}{2\sqrt{2}-2}\right)\left\|(\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top)\mathbf{Q}\mathbf{Q}^\top\right\|_F^2.$$

The proof relies on the following lemma.

Lemma C.3.2. *[125, Lemma E.1] Let* $\mathbf{U}$ *and* $\mathbf{Z}$ *be any two matrices in* $\mathbb{R}^{n\times r}$ *such that* $\mathbf{U}^\top\mathbf{Z} = \mathbf{Z}^\top\mathbf{U}$ *is PSD. Then*

$$\left\|(\mathbf{U}-\mathbf{Z})\,\mathbf{U}^\top\right\|_F^2 \le \frac{1}{2\sqrt{2}-2}\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top\right\|_F^2.$$

Proof of Lemma 4.3.4. Define two orthogonal projectors

$$\mathcal{Q} = \mathbf{Q}\mathbf{Q}^\top \qquad \text{and} \qquad \mathcal{Q}_\perp = \mathbf{Q}_\perp\mathbf{Q}_\perp^\top,$$

so $\mathcal{Q}$ is the orthogonal projector onto $\mathrm{Range}(\mathbf{U})$ and $\mathcal{Q}_\perp$ is the orthogonal projector onto the orthogonal complement of $\mathrm{Range}(\mathbf{U})$. Then

$$\begin{aligned}
\|(\mathbf{U}-\mathbf{Z})\mathbf{U}^\top\|_F^2 &\overset{①}{=} \|(\mathbf{U}-\mathcal{Q}\mathbf{Z})\mathbf{U}^\top\|_F^2 + \|\mathcal{Q}_\perp\mathbf{U}^\top\|_F^2 \\
&\overset{②}{=} \|(\mathbf{U}-\mathcal{Q}\mathbf{Z})\mathbf{U}^\top\|_F^2 + \langle\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}, \mathbf{U}^\top\mathbf{U}\rangle \\
&\overset{③}{\le} \frac{1}{2\sqrt{2}-2}\|\mathbf{U}\mathbf{U}^\top - (\mathcal{Q}\mathbf{Z})(\mathcal{Q}\mathbf{Z})^\top\|_F^2 + \langle\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}, \mathbf{U}^\top\mathbf{U} - \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\rangle + \langle\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}, \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\rangle \\
&\overset{④}{\le} \frac{1}{2\sqrt{2}-2}\|\mathbf{U}\mathbf{U}^\top - \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\|_F^2 + \langle\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}, \mathbf{U}^\top\mathbf{U} - \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\rangle + \langle\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}, \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\rangle \\
&\overset{⑤}{\le} \frac{1}{2\sqrt{2}-2}\|\mathbf{U}\mathbf{U}^\top - \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\|_F^2 + \frac{1}{8}\|\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}\|_F^2 + 2\|\mathbf{U}^\top\mathbf{U} - \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\|_F^2 + \langle\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}, \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\rangle,
\end{aligned} \tag{C.5}$$

where ① is by expressing $(\mathbf{U}-\mathbf{Z})\mathbf{U}^\top$ as the sum of two orthogonal factors $(\mathbf{U}-\mathcal{Q}\mathbf{Z})\mathbf{U}^\top$ and $-\mathcal{Q}_\perp\mathbf{Z}\mathbf{U}^\top$. ② is because $\|\mathcal{Q}_\perp\mathbf{Z}\mathbf{U}^\top\|_F^2 = \langle\mathcal{Q}_\perp\mathbf{Z}\mathbf{U}^\top, \mathcal{Q}_\perp\mathbf{Z}\mathbf{U}^\top\rangle = \langle\mathcal{Q}_\perp\mathbf{Z}\mathbf{U}^\top, \mathbf{Z}\mathbf{U}^\top\rangle = \langle\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}, \mathbf{U}^\top\mathbf{U}\rangle$. ③ uses Lemma C.3.2 by noting that $\mathbf{U}^\top\mathcal{Q}\mathbf{Z} = (\mathcal{Q}\mathbf{U})^\top\mathbf{Z} = \mathbf{U}^\top\mathbf{Z} \succeq 0$ satisfying the assumptions of Lemma C.3.2. ④ uses the fact that $\|\mathbf{U}\mathbf{U}^\top - (\mathcal{Q}\mathbf{Z})(\mathcal{Q}\mathbf{Z})^\top\|_F^2 = \|\mathbf{U}\mathbf{U}^\top - \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}\|_F^2 \le \|\mathbf{U}\mathbf{U}^\top - \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}\|_F^2 + \|\mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}_\perp\|_F^2 = \|\mathbf{U}\mathbf{U}^\top - \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q} - \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}_\perp\|_F^2 = \|\mathbf{U}\mathbf{U}^\top - \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\|_F^2$. ⑤ uses the following basic inequality that

$$\frac{1}{8}\|\mathbf{A}\|_F^2 + 2\|\mathbf{B}\|_F^2 \ge 2\sqrt{\frac{2}{8}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2} = \|\mathbf{A}\|_F\|\mathbf{B}\|_F \ge \langle\mathbf{A}, \mathbf{B}\rangle,$$

where $\mathbf{A} = \mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}$ and $\mathbf{B} = \mathbf{U}^\top\mathbf{U} - \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}$.

The Remaining Steps. The remaining steps involve showing the following bounds:

$$\|\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}\|_F^2 \le \|\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top\|_F^2, \tag{C.6}$$

$$\langle\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}, \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\rangle \le \|\mathbf{U}\mathbf{U}^\top - \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\|_F^2, \tag{C.7}$$

$$\|\mathbf{U}^\top\mathbf{U} - \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\|_F^2 \le \|\mathbf{U}\mathbf{U}^\top - \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\|_F^2. \tag{C.8}$$

This is because when plugging these bounds (C.6)- (C.8) into (C.5), we can obtain the desired result:

$$\|(\mathbf{U}-\mathbf{Z})\mathbf{U}^\top\|_F^2 \le \frac{1}{8}\|\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top\|_F^2 + \left(3 + \frac{1}{2\sqrt{2}-2}\right)\|(\mathbf{U}\mathbf{U}^\top - \mathbf{Z}\mathbf{Z}^\top)QQ^\top\|_F^2.$$

Showing (C.6).

$$\begin{aligned}\|\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}\|_F^2 &= \langle\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}_\perp, \mathcal{Q}_\perp\mathbf{Z}\mathbf{Z}^\top\rangle \\ &\overset{①}{=} \langle\mathcal{Q}_\perp\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}_\perp, \mathcal{Q}_\perp\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}_\perp\rangle \\ &= \|\mathcal{Q}_\perp\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}_\perp\|_F^2 \\ &\overset{②}{=} \|\mathcal{Q}_\perp(\mathbf{Z}\mathbf{Z}^\top - \mathbf{U}\mathbf{U}^\top)\mathcal{Q}_\perp\|_F^2 \\ &\overset{③}{\leq} \|\mathbf{Z}\mathbf{Z}^\top - \mathbf{U}\mathbf{U}^\top\|_F^2,\end{aligned}$$

where ① follows from the idempotence property that $\mathcal{Q}_\perp = \mathcal{Q}_\perp\mathcal{Q}_\perp$. ② follows from $\mathcal{Q}_\perp\mathbf{U} = \mathbf{0}$. ③ follows from the nonexpansiveness of projection operator: $\|\mathcal{Q}_\perp(\mathbf{Z}\mathbf{Z}^\top - \mathbf{U}\mathbf{U}^\top)\mathcal{Q}_\perp\|_F \leq \|(\mathbf{Z}\mathbf{Z}^\top - \mathbf{U}\mathbf{U}^\top)\mathcal{Q}_\perp\|_F \leq \|\mathbf{Z}\mathbf{Z}^\top - \mathbf{U}\mathbf{U}^\top\|_F$.

Showing Eq. (C.7). The argument here is pretty similar to that for (C.6):

$$\begin{aligned}\langle\mathbf{Z}^\top\mathcal{Q}_\perp\mathbf{Z}, \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\rangle &= \langle\mathcal{Q}\mathbf{Z}\mathbf{Z}^\top, \mathbf{Z}\mathbf{Z}^\top\mathcal{Q}_\perp\rangle \\ &= \langle\mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}_\perp, \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}_\perp\rangle \\ &= \|\mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}_\perp\|_F^2 \\ &\overset{①}{=} \|\mathcal{Q}(\mathbf{Z}\mathbf{Z}^\top - \mathbf{U}\mathbf{U}^\top)\mathcal{Q}_\perp\|_F^2 \\ &\overset{②}{\leq} \|\mathcal{Q}\mathbf{Z}\mathbf{Z}^\top - \mathbf{U}\mathbf{U}^\top\|_F^2,\end{aligned}$$

where ① is by $\mathcal{Q}_\perp\mathbf{U} = \mathbf{0}$. ② uses the nonexpansiveness of projection operator and $\mathcal{Q}\mathbf{U}\mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top$.

Showing Eq. (C.8). First by expanding $\|\cdot\|_F^2$ using inner products, (C.8) is equivalent to the following inequality

$$\|\mathbf{U}^\top\mathbf{U}\|_F^2 + \|\mathbf{U}^\top\mathbf{U} - \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\|_F^2 - 2\langle\mathbf{U}^\top\mathbf{U}, \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\rangle \leq \|\mathbf{U}\mathbf{U}^\top\|_F^2 + \|\mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\|_F^2 - 2\langle\mathbf{U}\mathbf{U}^\top, \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\rangle. \tag{C.9}$$

First of all, we recognize that

$$\begin{aligned}&\|\mathbf{U}^\top\mathbf{U}\|_F^2 = \sum_i \sigma_i(\mathbf{U})^2 = \|\mathbf{U}\mathbf{U}^\top\|_F^2; \\ &\|\mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\|_F^2 = \langle\mathbf{Z}^\top\mathcal{Q}\mathbf{Z}, \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\rangle = \langle\mathcal{Q}\mathbf{Z}\mathbf{Z}^\top, \mathbf{Z}\mathbf{Z}^\top\mathcal{Q}\rangle = \langle\mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}, \mathbf{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}\rangle = \|\mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}\|_F^2 \leq \|\mathbf{Z}\mathbf{Z}^\top\mathcal{Q}\|_F^2,\end{aligned}$$

where we use the idempotence and nonexpansiveness property of the projection matrix $\mathcal{Q}$ in the second line. Plugging these to (C.9), we find (C.9) reduces to

$$\langle\mathbf{U}^\top\mathbf{U}, \mathbf{Z}^\top\mathcal{Q}\mathbf{Z}\rangle \geq \langle\mathbf{U}\mathbf{U}^\top, \mathcal{Q}\mathbf{Z}\mathbf{Z}^\top\rangle = \langle\mathbf{U}\mathbf{U}^\top, \mathbf{Z}\mathbf{Z}^\top\rangle = \|\mathbf{U}^\top\mathbf{Z}\|_F^2. \tag{C.10}$$

To show (C.10), let $\mathbf{Q}\mathbf{\Sigma}\mathbf{P}^\top$ be the SVD of $\mathbf{U}$ with $\mathbf{\Sigma} \in \mathbb{R}^{r'\times r'}$ and $\mathbf{P} \in \mathbb{R}^{r\times r'}$ where r' is rank of $\mathbf{U}$. Then

$$\mathbf{U}^\top\mathbf{U} = \mathbf{P}\mathbf{\Sigma}^2\mathbf{P}^\top, \qquad \mathbf{Q} = \mathbf{U}\mathbf{P}\mathbf{\Sigma}^{-1} \qquad \text{and} \qquad \mathcal{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{U}\mathbf{P}\mathbf{\Sigma}^{-2}\mathbf{P}^\top\mathbf{U}^\top. \tag{C.11}$$

Now

$$\begin{aligned}
\text{LHS of } (C.10) &= \langle \mathbf{U}^\top \mathbf{U}, \mathbf{Z}^\top \mathcal{Q} \mathbf{Z} \rangle \\
&\overset{①}{=} \langle \mathbf{P}\mathbf{\Sigma}^2 \mathbf{P}^\top, \mathbf{Z}^\top U P \mathbf{\Sigma}^{-2} \mathbf{P}^\top \mathbf{U}^\top \mathbf{Z} \rangle \\
&\overset{②}{=} \langle \mathbf{\Sigma}^2, \mathbf{P}^\top (\mathbf{U}^\top \mathbf{Z}) P \mathbf{\Sigma}^{-2} \mathbf{P}^\top (\mathbf{U}^\top \mathbf{Z}) \mathbf{P} \rangle \\
&\overset{③}{=} \langle \mathbf{\Sigma}^2, \mathbf{G}\mathbf{\Sigma}^{-2}\mathbf{G} \rangle \\
&= \|\mathbf{\Sigma}\mathbf{G}\mathbf{\Sigma}^{-1}\|_F^2 \\
&\overset{④}{\geq} \|\mathbf{G}\|_F^2 \\
&\overset{⑤}{=} \|\mathbf{U}^\top \mathbf{Z}\|_F^2,
\end{aligned}$$

where ① is by (C.11) and ② uses the assumption that $\mathbf{Z}^\top \mathbf{U} = \mathbf{U}^\top \mathbf{Z} \succeq 0$. In ③, we define $\mathbf{G} := \mathbf{P}^\top (\mathbf{U}^\top \mathbf{Z}) \mathbf{P}$. ⑤ is because $\|\mathbf{G}\|_F^2 = \|\mathbf{P}^\top (\mathbf{U}^\top \mathbf{Z}) \mathbf{P}\|_F^2 = \|\mathbf{U}^\top \mathbf{Z}\|_F^2$ due to the rotational invariance of $\|\cdot\|_F$. ④ is because

$$\begin{aligned}
\|\mathbf{\Sigma}\mathbf{G}\mathbf{\Sigma}^{-1}\|_F^2 &= \sum_{i,j} \frac{\sigma_i^2}{\sigma_j^2} G_{ij}^2 \\
&= \sum_{i=j} G_{ii}^2 + \sum_{i>j} \left(\frac{\sigma_i^2}{\sigma_j^2} + \frac{\sigma_j^2}{\sigma_i^2} \right) G_{ij}^2 \\
&\geq \sum_{i=j} G_{ii}^2 + \sum_{i>j} 2 \left(\frac{\sigma_i}{\sigma_j} \right) \left(\frac{\sigma_j}{\sigma_i} \right) G_{ij}^2 \\
&= \sum_{i,j} G_{ij}^2 \\
&= \|\mathbf{G}\|_F^2,
\end{aligned}$$

where the second line follows from the symmetric property of $\mathbf{G}$ since $\mathbf{G} = \mathbf{P}^\top (\mathbf{U}^\top \mathbf{Z}) \mathbf{P} \succeq 0$ and $\mathbf{U}^\top \mathbf{Z} \succeq 0$. □

C.4 Proof of Lemma 4.3.5

Lemma C.4.1 (Lemma 4.3.5)**.** *Suppose the objective function $f(\mathbf{X})$ in $(\mathcal{P}_0)$ is twice continuously differentiable and satisfies the restricted well-conditionedness assumption $(\mathcal{C})$. Further, let $\mathbf{U}$ be any critical point of $(\mathcal{F}_0)$ and $\mathbf{Q}$ be the orthonormal basis spanning* $\mathrm{Range}(\mathbf{U})$*. Then*

$$\left\| (\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star \mathbf{U}^{\star\top}) \mathbf{Q}\mathbf{Q}^\top \right\|_F \leq \frac{\beta - \alpha}{\beta + \alpha} \left\| \mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star \mathbf{U}^{\star\top} \right\|_F.$$

Proof. Let $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$ and $\mathbf{X}^\star = \mathbf{U}^\star \mathbf{U}^{\star\top}$. We start with the critical point condition $\nabla f(\mathbf{X})\mathbf{U} = \mathbf{0}$ which implies

$$\nabla f(\mathbf{X}) \mathbf{U}\mathbf{U}^\dagger = \nabla f(\mathbf{X}) \mathbf{Q}\mathbf{Q}^\top = \mathbf{0},$$

where † denotes the pseudoinverse. Then for all $Z \in \mathbb{R}^{n \times n}$, we have

$$\Rightarrow \langle \nabla f(\mathbf{X}), \mathbf{Z}\mathbf{Q}\mathbf{Q}^\top \rangle = 0$$

$$
\begin{aligned}
&\overset{①}{\Rightarrow} \langle \nabla f(\mathbf{X}^\star) + \int_0^1 [\nabla^2 f(t\mathbf{X} + (1-t)\mathbf{X}^\star)](\mathbf{X} - \mathbf{X}^\star)dt, \mathbf{Z}\mathbf{Q}\mathbf{Q}^\top \rangle = 0 \\
&\Rightarrow \langle \nabla f(\mathbf{X}^\star), \mathbf{Z}\mathbf{Q}\mathbf{Q}^\top \rangle + \left[\int_0^1 \nabla^2 f(t\mathbf{X} + (1-t)\mathbf{X}^\star)dt\right](\mathbf{X} - \mathbf{X}^\star, \mathbf{Z}\mathbf{Q}\mathbf{Q}^\top) = 0 \\
&\overset{②}{\Rightarrow} \left| -\frac{2}{\beta+\alpha}\langle \nabla f(\mathbf{X}^\star), \mathbf{Z}\mathbf{Q}\mathbf{Q}^\top \rangle - \langle \mathbf{X} - \mathbf{X}^\star, \mathbf{Z}\mathbf{Q}\mathbf{Q}^\top \rangle \right| \leq \frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{X} - \mathbf{X}^\star\|_F \|Z\mathbf{Q}\mathbf{Q}^\top\|_F \\
&\Rightarrow \left| \frac{2}{\beta+\alpha}\langle \nabla f(\mathbf{X}^\star), \mathbf{Z}\mathbf{Q}\mathbf{Q}^\top \rangle + \langle \mathbf{X} - \mathbf{X}^\star, \mathbf{Z}\mathbf{Q}\mathbf{Q}^\top \rangle \right| \leq \frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{X} - \mathbf{X}^\star\|_F \|\mathbf{Z}\mathbf{Q}\mathbf{Q}^\top\|_F \\
&\overset{③}{\Rightarrow} \left| \frac{2}{\beta+\alpha}\langle \nabla f(\mathbf{X}^\star), (\mathbf{X} - \mathbf{X}^\star)\mathbf{Q}\mathbf{Q}^\top \rangle + \|(\mathbf{X} - \mathbf{X}^\star)\mathbf{Q}\mathbf{Q}^\top\|_F^2 \right| \leq \frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{X} - \mathbf{X}^\star\|_F \|(\mathbf{X} - \mathbf{X}^\star)\mathbf{Q}\mathbf{Q}^\top\|_F \\
&\overset{④}{\Rightarrow} \frac{2}{\beta+\alpha}\langle \nabla f(\mathbf{X}^\star), (\mathbf{X} - \mathbf{X}^\star)\mathbf{Q}\mathbf{Q}^\top \rangle + \|(\mathbf{X} - \mathbf{X}^\star)\mathbf{Q}\mathbf{Q}^\top\|_F^2 \leq \frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{X} - \mathbf{X}^\star\|_F \|(\mathbf{X} - \mathbf{X}^\star)\mathbf{Q}\mathbf{Q}^\top\|_F \\
&\Rightarrow \|(\mathbf{X} - \mathbf{X}^\star)\mathbf{Q}\mathbf{Q}^\top\|_F \leq \delta\|\mathbf{X} - \mathbf{X}^\star\|_F,
\end{aligned}
$$

where ① uses the Taylor's Theorem for vector-valued functions [148, Eq. (2.5) in Theorem 2.1]. ② uses Proposition 4.2.1 by noting that the PSD matrix $[t\mathbf{X}^\star + (1-t)\mathbf{X}]$ has rank at most $2r$ for all $t \in [0,1]$ and $\mathrm{rank}(\mathbf{X} - \mathbf{X}^\star) \leq 4r, \mathrm{rank}(\mathbf{Z}\mathbf{Q}\mathbf{Q}^\top) \leq 4r$. ③ is by choosing $\mathbf{Z} = X - \mathbf{X}^\star$. ④ follows from $\langle \nabla f(\mathbf{X}^\star), (\mathbf{X} - \mathbf{X}^\star)\mathbf{Q}\mathbf{Q}^\top \rangle \geq 0$ since

$$
\langle \nabla f(\mathbf{X}^\star), (\mathbf{X} - \mathbf{X}^\star)\mathbf{Q}\mathbf{Q}^\top \rangle \overset{\text{(i)}}{=} \langle \nabla f(\mathbf{X}^\star), \mathbf{X} - \mathbf{X}^\star\mathbf{Q}\mathbf{Q}^\top \rangle \overset{\text{(ii)}}{=} \langle \nabla f(\mathbf{X}^\star), \mathbf{X} \rangle \overset{\text{(iii)}}{\geq} 0,
$$

where (i) follows from $X\mathbf{Q}\mathbf{Q}^\top = \mathbf{U}\mathbf{U}^\top\mathbf{Q}\mathbf{Q}^\top = \mathbf{U}\mathbf{U}^\top$ since $\mathbf{Q}\mathbf{Q}^\top$ is the orthogonal projector onto $\mathrm{Range}(\mathbf{U})$. (ii) uses the fact that

$$
\nabla f(\mathbf{X}^\star)\mathbf{X}^\star = \mathbf{0} = \mathbf{X}^\star\nabla f(\mathbf{X}^\star),
$$

and (iii) is because $\nabla f(\mathbf{X}^\star) \succeq 0, \mathbf{X} \succeq 0$. □

C.5 Proof of Proposition 4.4.1

Proposition C.5.1 (Proposition 4.4.1)**.** *Any critical point* $(\mathbf{U}, \mathbf{V}) \in \mathcal{X}$ *forms a balanced pair in* $\mathcal{E}$.

For any critical point $(\mathbf{U}, \mathbf{V})$, we have

$$
\nabla g(\mathbf{U}, \mathbf{V}) = \Xi(\mathbf{U}\mathbf{V}^\top)\mathbf{W} = \mathbf{0},
$$

where $\mathbf{W} = \begin{bmatrix} \mathbf{U}^\top & \mathbf{V}^\top \end{bmatrix}^\top$. Further denote $\widehat{\mathbf{W}} = \begin{bmatrix} \mathbf{U}^\top & -\mathbf{V}^\top \end{bmatrix}^\top$. Then

$$
\begin{aligned}
&\overset{①}{\Rightarrow}\widehat{\mathbf{W}}^\top \nabla g(\mathbf{U},\mathbf{V}) + \nabla g(\mathbf{U},\mathbf{V})^\top \widehat{\mathbf{W}} = \mathbf{0} \\
&\overset{②}{\Rightarrow}\widehat{\mathbf{W}}^\top \Xi(\mathbf{U}\mathbf{V}^\top)\mathbf{W} + \mathbf{W}^\top \Xi(\mathbf{U}\mathbf{V}^\top)\widehat{\mathbf{W}} = \mathbf{0} \\
&\overset{③}{\Rightarrow}[\mathbf{U}^\top \ -\mathbf{V}^\top] \begin{bmatrix} \lambda \mathbf{I} & \nabla f(\mathbf{U}\mathbf{V}^\top) \\ \nabla f(\mathbf{U}\mathbf{V}^\top)^\top & \lambda \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} + [\mathbf{U}^\top \ \mathbf{V}^\top] \begin{bmatrix} \lambda \mathbf{I} & \nabla f(\mathbf{U}\mathbf{V}^\top) \\ \nabla f(\mathbf{U}\mathbf{V}^\top)^\top & \lambda \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{U} \\ -\mathbf{V} \end{bmatrix} = \mathbf{0} \\
&\overset{④}{\Rightarrow}\lambda \left(2\mathbf{U}^\top \mathbf{U} - 2\mathbf{V}^\top \mathbf{V}\right) + \underbrace{\mathbf{U}^\top \left(\nabla f(\mathbf{U}\mathbf{V}^\top) - \nabla f(\mathbf{U}\mathbf{V}^\top)\right)\mathbf{V}}_{=\mathbf{0}} + \underbrace{\mathbf{V}^\top \left(\nabla f(\mathbf{U}\mathbf{V}^\top)^\top - \nabla f(\mathbf{U}\mathbf{V}^\top)^\top\right)\mathbf{U}}_{=\mathbf{0}} = \mathbf{0} \\
&\Rightarrow 2\lambda(\mathbf{U}^\top \mathbf{U} - \mathbf{V}^\top \mathbf{V}) = \mathbf{0} \\
&\overset{⑤}{\Rightarrow}\mathbf{U}^\top \mathbf{U} - \mathbf{V}^\top \mathbf{V} = \mathbf{0},
\end{aligned}
$$

where ① follows from $\nabla g(\mathbf{U},\mathbf{V}) = \mathbf{0}$ and ② follows from $\nabla g(\mathbf{U},\mathbf{V}) = \Xi(\mathbf{U}\mathbf{V}^\top)W$. ③ follows by plugging the definitions of $W, \widehat{\mathbf{W}}$ and $\Xi(\cdot)$ into the second line. ④ follows from direct computations. ⑤ holds since $\lambda > 0$. □

C.6 Proof of Lemma 4.4.1

Lemma C.6.1 (Lemma 4.4.1)**.** *Let* $\mathbf{W} = \begin{bmatrix} \mathbf{U}^\top & \mathbf{V}^\top \end{bmatrix}^\top$ *with* $(\mathbf{U},\mathbf{V}) \in \mathcal{E}$. *Then for every* $\mathbf{D} = \begin{bmatrix} \mathbf{D}_\mathbf{U}^\top & \mathbf{D}_\mathbf{V}^\top \end{bmatrix}^\top$ *of proper dimension, we have*

$$\|\mathcal{P}_{\mathrm{on}}(\mathbf{D}\mathbf{W}^\top)\|_F^2 = \|\mathcal{P}_{\mathrm{off}}(\mathbf{D}\mathbf{W}^\top)\|_F^2.$$

First recall

$$\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}, \qquad \widehat{\mathbf{W}} = \begin{bmatrix} \mathbf{U} \\ -\mathbf{V} \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} \mathbf{D}_\mathbf{U} \\ \mathbf{D}_\mathbf{V} \end{bmatrix}, \qquad \widehat{\mathbf{D}} = \begin{bmatrix} \mathbf{D}_\mathbf{U} \\ -\mathbf{D}_\mathbf{V} \end{bmatrix}.$$

By performing the following change of variables

$$\mathbf{W}_1 \leftarrow \mathbf{D}, \qquad \widehat{\mathbf{W}}_1 \leftarrow \widehat{\mathbf{D}}, \qquad \mathbf{W}_2 \leftarrow \mathbf{W}, \qquad \widehat{\mathbf{W}}_2 \leftarrow \widehat{\mathbf{W}}$$

in (4.29), we have

$$
\begin{aligned}
\|\mathcal{P}_{\mathrm{on}}(\mathbf{D}\mathbf{W}^\top)\|_F^2 &= \frac{1}{4}\|\mathbf{D}\mathbf{W}^\top + \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top\|_F^2 = \frac{1}{4}\langle \mathbf{D}\mathbf{W}^\top + \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top, \mathbf{D}\mathbf{W}^\top + \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top \rangle; \\
\|\mathcal{P}_{\mathrm{off}}(\mathbf{D}\mathbf{W}^\top)\|_F^2 &= \frac{1}{4}\|\mathbf{D}\mathbf{W}^\top - \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top\|_F^2 = \frac{1}{4}\langle \mathbf{D}\mathbf{W}^\top - \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top, \mathbf{D}\mathbf{W}^\top - \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top \rangle.
\end{aligned}
$$

Then it implies that

$$
\begin{aligned}
\|\mathcal{P}_{\mathrm{on}}(\mathbf{D}\mathbf{W}^\top)\|_F^2 - \|\mathcal{P}_{\mathrm{off}}(\mathbf{D}\mathbf{W}^\top)\|_F^2 &= \frac{1}{4}\langle \mathbf{D}\mathbf{W}^\top + \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top, \mathbf{D}\mathbf{W}^\top + \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top \rangle - \frac{1}{4}\langle \mathbf{D}\mathbf{W}^\top - \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top, \mathbf{D}\mathbf{W}^\top - \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top \rangle \\
&= \langle \mathbf{D}\mathbf{W}^\top, \widehat{\mathbf{D}}\widehat{\mathbf{W}}^\top \rangle = \langle \widehat{\mathbf{D}}^\top \mathbf{D}, \widehat{\mathbf{W}}^\top \mathbf{W} \rangle = 0,
\end{aligned}
$$

since $\widehat{\mathbf{W}}^\top W = \mathbf{0}$ from (4.27). □

C.7 Proof of Lemma 4.4.2

Lemma C.7.1 (Lemma 4.4.2)**.** *Let* $\mathbf{W}_1 = \begin{bmatrix} \mathbf{U}_1^\top & \mathbf{V}_1^\top \end{bmatrix}^\top$, $\mathbf{W}_2 = \begin{bmatrix} \mathbf{U}_2^\top & \mathbf{V}_2^\top \end{bmatrix}^\top$ *with* $(\mathbf{U}_1, \mathbf{V}_1), (\mathbf{U}_2, \mathbf{V}_2) \in \mathcal{E}$*. Then*

$$\|\mathcal{P}_{\mathrm{on}}(\mathbf{W}_1\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{W}_2^\top)\|_F^2 \le \|\mathcal{P}_{\mathrm{off}}(\mathbf{W}_1\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{W}_2^\top)\|_F^2.$$

Proof. To begin with, we define $\widehat{\mathbf{W}}_1 = \begin{bmatrix} \mathbf{U}_1 \\ -\mathbf{V}_1 \end{bmatrix}$, $\widehat{\mathbf{W}}_2 = \begin{bmatrix} \mathbf{U}_2 \\ -\mathbf{V}_2 \end{bmatrix}$. Then

$$\begin{aligned}
&\|\mathcal{P}_{\mathrm{on}}(\mathbf{W}_1\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{W}_2^\top)\|_F^2 - \|\mathcal{P}_{\mathrm{off}}(\mathbf{W}_1\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{W}_2^\top)\|_F^2 \\
&\overset{①}{=} \|\mathcal{P}_{\mathrm{on}}(\mathbf{W}_1\mathbf{W}_1^\top) - \mathcal{P}_{\mathrm{on}}(\mathbf{W}_2\mathbf{W}_2^\top)\|_F^2 - \|\mathcal{P}_{\mathrm{off}}(\mathbf{W}_1\mathbf{W}_1^\top) - \mathcal{P}_{\mathrm{off}}(\mathbf{W}_2\mathbf{W}_2^\top)\|_F^2 \\
&\overset{②}{=} \left\| \frac{\mathbf{W}_1\mathbf{W}_1^\top + \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^\top}{2} - \frac{\mathbf{W}_2\mathbf{W}_2^\top + \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^\top}{2} \right\|_F^2 - \left\| \frac{\mathbf{W}_1\mathbf{W}_1^\top - \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^\top}{2} - \frac{\mathbf{W}_2\mathbf{W}_2^\top - \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^\top}{2} \right\|_F^2 \\
&= \left\| \frac{\mathbf{W}_1\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{W}_2^\top}{2} + \frac{\widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^\top - \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^\top}{2} \right\|_F^2 - \left\| \frac{\mathbf{W}_1\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{W}_2^\top}{2} - \frac{\widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^\top - \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^\top}{2} \right\|_F^2 \\
&\overset{③}{=} \langle \mathbf{W}_1\mathbf{W}_1^\top - \mathbf{W}_2\mathbf{W}_2^\top, \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^\top - \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^\top \rangle \\
&= \langle \mathbf{W}_1\mathbf{W}_1^\top, \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^\top \rangle + \langle \mathbf{W}_2\mathbf{W}_2^\top, \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^\top \rangle - \langle \mathbf{W}_1\mathbf{W}_1^\top, \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^\top \rangle - \langle \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^\top, \mathbf{W}_2\mathbf{W}_2^\top \rangle \\
&\overset{④}{=} -\langle \mathbf{W}_1\mathbf{W}_1^\top, \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^\top \rangle - \langle \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^\top, \mathbf{W}_2\mathbf{W}_2^\top \rangle \\
&\overset{⑤}{\le} 0,
\end{aligned}$$

where ① is due to the linearity of $\mathcal{P}_{\mathrm{on}}$ and $\mathcal{P}_{\mathrm{off}}$. ② follows from (4.29). ③ is by expanding $\|\cdot\|_F^2$. ④ comes from (4.27) that

$$\widehat{\mathbf{W}}_i^\top \mathbf{W}_i = \mathbf{W}_i^\top \widehat{\mathbf{W}}_i = \mathbf{0}, \qquad \text{for } i = 1, 2.$$

⑤ uses the fact that

$$\mathbf{W}_1\mathbf{W}_1^\top \succeq 0, \qquad \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^\top \succeq 0, \qquad \mathbf{W}_2\mathbf{W}_2^\top \succeq 0, \qquad \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^\top \succeq 0.$$

□

C.8 Proof of Proposition 4.4.2

Proposition C.8.1 (Proposition 4.4.2)**.** *Any* $(\mathbf{U}^\star, \mathbf{V}^\star)$ *in* (4.22) *is a global optimum of the factored program* $(\mathcal{F}_1)$*:*

$$g(\mathbf{U}^\star, \mathbf{V}^\star) \le g(\mathbf{U}, \mathbf{V}), \text{ for all } \mathbf{U} \in \mathbb{R}^{n\times r}, \mathbf{V} \in \mathbb{R}^{m\times r}.$$

Proof. From (4.22), we have

$$\begin{aligned}\frac{1}{2}\left(\|\mathbf{U}^\star\|_F^2+\|\mathbf{V}^\star\|_F^2\right) &\overset{①}{=} \frac{1}{2}\left(\left\|\mathbf{P}^\star[\sqrt{\mathbf{\Sigma}^\star}\ \mathbf{0}_{r^\star\times(r-r^\star)}]\mathbf{R}\right\|_F^2+\left\|\mathbf{Q}^\star[\sqrt{\mathbf{\Sigma}^\star}\ \mathbf{0}_{r^\star\times(r-r^\star)}]\mathbf{R}\right\|_F^2\right)\\ &\overset{②}{=} \frac{1}{2}\left(\left\|\sqrt{\mathbf{\Sigma}^\star}\right\|_F^2+\left\|\sqrt{\mathbf{\Sigma}^\star}\right\|_F^2\right)\\ &= \left\|\sqrt{\mathbf{\Sigma}^\star}\right\|_F^2\\ &\overset{③}{=} \|\mathbf{X}^\star\|_*,\end{aligned}$$

where ① uses the definitions of $\mathbf{U}^\star$ and $\mathbf{V}^\star$ in (4.22). ② uses the rotational invariance of $\|\cdot\|_F$. ③ is because $\|\sqrt{\mathbf{\Sigma}^\star}\|_F^2=\sum_j\sigma_k(\mathbf{X}^\star)=\|\mathbf{X}^\star\|_*$.

Therefore,

$$\begin{aligned}f(\mathbf{U}^\star\mathbf{V}^{\star\top})+\lambda(\|\mathbf{U}^\star\|_F^2+\|V^\star\|_F^2)/2 &\overset{①}{=} f(\mathbf{X}^\star)+\lambda\|\mathbf{X}^\star\|_*\\ &\le f(\mathbf{X})+\lambda\|\mathbf{X}\|_*\\ &\overset{②}{=} f(\mathbf{U}\mathbf{V}^\top)+\lambda\|\mathbf{U}\mathbf{V}^\top\|_*\\ &\overset{③}{\le} f(\mathbf{U}\mathbf{V}^\top)+\lambda(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2)/2,\end{aligned}$$

where ① comes from the optimality of $\mathbf{X}^\star$ for $(\mathcal{P}_1)$. ② is by choosing $\mathbf{X}=\mathbf{U}\mathbf{V}^\top$. ③ is because $\|\mathbf{U}\mathbf{V}^\top\|_*\le(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2)/2$ by the optimization formulation of the matrix nuclear norm [65, Lemma 5.1] that

$$\|\mathbf{X}\|_*=\min_{\mathbf{X}=\mathbf{U}\mathbf{V}^\top}\frac{1}{2}(\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2).$$

□

C.9 Proof of Lemma 4.4.3

Lemma C.9.1 (Lemma 4.4.3)**.** *Suppose the function $f(\mathbf{X})$ in $(\mathcal{P}_1)$ is restricted well-conditioned $(\mathcal{C})$. Let $\mathbf{W}=\begin{bmatrix}\mathbf{U}^\top & \mathbf{V}^\top\end{bmatrix}^\top$ with $(\mathbf{U},\mathbf{V})\in\mathcal{X}$, $\mathbf{W}^\star=\begin{bmatrix}\mathbf{U}^{\star\top} & \mathbf{V}^{\star\top}\end{bmatrix}^\top$ correspond to the global optimum of $(\mathcal{P}_1)$ and $\mathbf{Q}\mathbf{Q}^\top$ be the orthogonal projector onto* $\mathrm{Range}(\mathbf{W})$*. Then*

$$\|(\mathbf{W}\mathbf{W}^\top-\mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\|_F\le 2\frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{U}\mathbf{V}^\top-\mathbf{X}^\star\|_F.$$

Proof. Let $\mathbf{Z}=\begin{bmatrix}\mathbf{Z}_\mathbf{U}\\ \mathbf{Z}_\mathbf{V}\end{bmatrix}$ with arbitrary $\mathbf{Z}_\mathbf{U}\in\mathbb{R}^{n\times r}$ and $\mathbf{Z}_\mathbf{V}\in\mathbb{R}^{m\times r}$. Then

$$\begin{aligned}&\Rightarrow\langle\Xi(\mathbf{X})\mathbf{W},\mathbf{Z}\rangle=\langle\mathbf{0},\mathbf{Z}\rangle=0\\ &\Rightarrow\left\langle\Xi(\mathbf{X})-\Xi(\mathbf{X}^\star)+\Xi(\mathbf{X}^\star),\mathbf{Z}\mathbf{W}^\top\right\rangle=0\\ &\Rightarrow\left\langle\begin{bmatrix}\lambda\mathbf{I} & \nabla f(\mathbf{X})\\ \nabla f(\mathbf{X})^\top & \lambda\mathbf{I}\end{bmatrix}-\begin{bmatrix}\lambda\mathbf{I} & \nabla f(\mathbf{X}^\star)\\ \nabla f(\mathbf{X}^\star)^\top & \lambda\mathbf{I}\end{bmatrix}+\Xi(\mathbf{X}^\star),\mathbf{Z}\mathbf{W}^\top\right\rangle=0\end{aligned}$$

$$
\begin{aligned}
&\Rightarrow \left\langle \begin{bmatrix} \mathbf{0} & \nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^\star) \\ \nabla f(\mathbf{X})^\top - \nabla f(\mathbf{X}^\star)^\top & \mathbf{0} \end{bmatrix} + \Xi(\mathbf{X}^\star), \mathbf{Z}\mathbf{W}^\top \right\rangle = 0 \\
&\Rightarrow \left\langle \begin{bmatrix} \mathbf{0} & \int_0^1 [\nabla^2 f(\mathbf{X}^\star + t(\mathbf{X} - \mathbf{X}^\star))](\mathbf{X} - \mathbf{X}^\star) dt \\ * & \mathbf{0} \end{bmatrix} + \Xi(\mathbf{X}^\star), \mathbf{Z}\mathbf{W}^\top \right\rangle = 0 \\
&\Rightarrow \left\langle \begin{bmatrix} \mathbf{0} & \int_0^1 [\nabla^2 f(\mathbf{X}^\star + t(\mathbf{X} - \mathbf{X}^\star))](\mathbf{X} - \mathbf{X}^\star) dt \\ * & \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{Z}_\mathbf{U}\mathbf{U}^\top & \mathbf{Z}_\mathbf{U}\mathbf{V}^\top \\ \mathbf{Z}_\mathbf{V}\mathbf{U}^\top & \mathbf{Z}_\mathbf{V}\mathbf{V}^\top \end{bmatrix} \right\rangle + \left\langle \Xi(\mathbf{X}^\star), \mathbf{Z}\mathbf{W}^\top \right\rangle = 0 \\
&\Rightarrow \int_0^1 [\nabla^2 f(\mathbf{X}^\star + t(\mathbf{X} - \mathbf{X}^\star))](\mathbf{X} - \mathbf{X}^\star, \mathbf{Z}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{Z}_\mathbf{V}^\top) dt + \left\langle \Xi(\mathbf{X}^\star), \mathbf{Z}\mathbf{W}^\top \right\rangle = 0,
\end{aligned}
$$

where the fifth line follows from the Taylor's Theorem for vector-valued functions [148, Eq. (2.5) in Theorem 2.1] and for convenience $* = \left(\int_0^1 [\nabla^2 f(\mathbf{X}^\star + t(\mathbf{X} - \mathbf{X}^\star))](\mathbf{X} - \mathbf{X}^\star) dt\right)^\top$ in the fifth and sixth lines. Then, from Proposition 4.2.1 and Eq. (4.29), we have

$$
\left| \frac{2}{\beta + \alpha} \underbrace{\left\langle \Xi(\mathbf{X}^\star), \mathbf{Z}\mathbf{W}^\top \right\rangle}_{\Pi_1(\mathbf{Z})} + \underbrace{\left\langle \mathcal{P}_{\text{off}}(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}), \mathbf{Z}\mathbf{W}^\top \right\rangle}_{\Pi_2(\mathbf{Z})} \right| \leq \frac{\beta - \alpha}{\beta + \alpha} \|\mathbf{X} - \mathbf{X}^\star\|_F \underbrace{\|\mathcal{P}_{\text{off}}(\mathbf{Z}\mathbf{W}^\top)\|_F}_{\Pi_3(\mathbf{Z})}. \tag{C.12}
$$

The Remaining Steps. The remaining steps are choosing $\mathbf{Z} = (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{W}^{\top\dagger}$ and showing the following

$$
\Pi_1(\mathbf{Z}) \geq 0, \tag{C.13}
$$

$$
\Pi_2(\mathbf{Z}) \geq \frac{1}{2}\|(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\|_F^2, \tag{C.14}
$$

$$
\Pi_3(\mathbf{Z}) \leq \|(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\|_F. \tag{C.15}
$$

Then plugging (C.13)- (C.15) into (C.12) yields the desired result:

$$
\frac{1}{2}\left\|(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\right\|_F^2 \leq \frac{\beta - \alpha}{\beta + \alpha}\|\mathbf{X} - \mathbf{X}^\star\|_F \left\|(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\right\|_F,
$$

or equivalently,

$$
\left\|(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\right\|_F \leq 2\frac{\beta - \alpha}{\beta + \alpha}\|\mathbf{X} - \mathbf{X}^\star\|_F.
$$

Showing (C.13). Choosing $\mathbf{Z} = (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{W}^{\top\dagger}$ and noting that $\mathbf{Q}\mathbf{Q}^\top = \mathbf{W}^T\mathbf{W}^{\top\dagger}$, we have $\mathbf{Z}\mathbf{W}^\top = (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{W}^{\top\dagger}\mathbf{W}^\top = (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top$. Then

$$
\Pi_1(\mathbf{Z}) = \langle \Xi(\mathbf{X}^\star), (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top \rangle = \langle \Xi(\mathbf{X}^\star), \mathbf{W}\mathbf{W}^\top \rangle \geq 0,
$$

where the second equality holds since $\mathbf{W}\mathbf{W}^\top\mathbf{Q}\mathbf{Q}^\top = \mathbf{W}\mathbf{W}^\top$ and $\Xi(\mathbf{X}^\star)\mathbf{W}^\star = \mathbf{0}$ by (4.25). The inequality is due to $\Xi(\mathbf{X}^\star) \succeq 0$.

Showing (C.14). First recognize that $\mathcal{P}_{\text{off}}(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}) = \frac{1}{2}(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top} - \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top + \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top})$. Then

$$\begin{aligned}\Pi_2(\mathbf{Z}) &= \langle \mathcal{P}_{\text{off}}(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}), \mathbf{Z}\mathbf{W}^\top \rangle \\ &= \frac{1}{2}\left\langle \mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}, (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top \right\rangle - \frac{1}{2}\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top - \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top \right\rangle.\end{aligned}$$

Therefore, (C.14) follows from

$$\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top - \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top \right\rangle = \left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, -\mathbf{W}^\star\mathbf{W}^{\star\top} \right\rangle + \left\langle -\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \mathbf{W}\mathbf{W}^\top \right\rangle \le 0,$$

where the first equality uses (4.27) and the inequality is because

$$\widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top \succeq 0, \qquad \mathbf{W}^\star\mathbf{W}^{\star\top} \succeq 0, \qquad \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top} \succeq 0, \qquad \mathbf{W}\mathbf{W}^\top \succeq 0.$$

Showing (C.15). Plugging $\mathbf{Z} = (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{W}^{\top\dagger}$ gives

$$\Pi_3(\mathbf{Z}) = \|\mathcal{P}_{\text{off}}((\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top)\|_F,$$

which is obviously no larger than $\|(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{Q}\mathbf{Q}^\top\|_F$ by the definition of the operation $\mathcal{P}_{\text{off}}$.

□

APPENDIX D

APPENDICES FOR CHAPTER 5

D.1 Proof of Lemma 5.3.1

Lemma D.1.1 (Lemma 5.3.1)**.** *Suppose* $\Omega = [n] \times [m]$*. Let*

$$\alpha_{q,\gamma} = \min_{|x| \le \gamma} \min \left(\frac{(q'(x))^2 - q(x)q''(x)}{q^2(x)}, \frac{(q'(x))^2 + (1-q(x))q''(x)}{(1-q(x))^2} \right)$$

and

$$\beta_{q,\gamma} = \max_{|x| \le \gamma} \max \left(\frac{(q'(x))^2 - q(x)q''(x)}{q^2(x)}, \frac{(q'(x))^2 + (1-q(x))q''(x)}{(1-q(x))^2} \right).$$

Then $F_{\Omega,\mathbf{Y}}$ *satisfies the restricted strong convexity and smoothness condition:*

$$\alpha_{q,\gamma} \|\mathbf{G}\|_F^2 \le [\nabla^2 F_{\Omega,\mathbf{Y}}(\mathbf{X})](\mathbf{G},\mathbf{G}) \le \beta_{q,\gamma} \|\mathbf{G}\|_F^2$$

for any $\mathbf{G} \in \mathbb{R}^{n \times m}$ *and* $\|\mathbf{X}\|_\infty \le \gamma$*.*

Proof of Lemma 5.3.1. We compute the partial derivative of $F_{\Omega,\mathbf{Y}}$ in terms of $X_{i,j}$ as

$$\frac{\partial F_{\Omega,\mathbf{Y}}}{\partial X_{i,j}} = -\mathbb{1}_{(Y_{i,j}=1)} \frac{q'(X_{i,j})}{q(X_{i,j})} + \mathbb{1}_{(Y_{i,j}=-1)} \frac{q'(X_{i,j})}{1-q(X_{i,j})},$$

which implies

$$\frac{\partial^2 F_{\Omega,\mathbf{Y}}}{\partial X_{i,j} \partial X_{i,j}} = \mathbb{1}_{(Y_{i,j}=1)} \frac{(q'(X_{i,j}))^2 - q(X_{i,j})q''(X_{i,j})}{q^2(X_{i,j})} + \mathbb{1}_{(Y_{i,j}=-1)} \frac{(q'(X_{i,j}))^2 + (1-q(X_{i,j}))q''(X_{i,j})}{(1-q(X_{i,j}))^2}$$

and

$$\frac{\partial^2 F_{\Omega,\mathbf{Y}}}{\partial X_{i,j} \partial X_{k,\ell}} = 0$$

for all $(k,\ell) \neq (i,j)$. Thus, the bilinear form for the Hessian of $\nabla^2 F_{\Omega,\mathbf{Y}}(\mathbf{X})$ can be computed as

$$[\nabla^2 F_{\Omega,\mathbf{Y}}(\mathbf{X})](\mathbf{G},\mathbf{G}) = \sum_i \sum_j \frac{\partial^2 F_{\Omega,\mathbf{Y}}}{\partial X_{i,j} \partial X_{i,j}} G_{i,j}^2$$

for any $\mathbf{G} \in \mathbb{R}^{n \times m}$. Now since by assumption $\|\mathbf{X}\|_\infty \le \gamma$, we have

$$\alpha_{q,\gamma} \|\mathbf{G}\|_F^2 \le [\nabla^2 F_{\Omega,\mathbf{Y}}(\mathbf{X})](\mathbf{G},\mathbf{G}) \le \beta_{q,\gamma} \|\mathbf{G}\|_F^2.$$

□

D.2 Proof of Proposition 5.4.1

Proposition D.2.1 (Proposition 5.4.1)**.** *Suppose the function $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (5.3) *with positive α and β. Then for any $n \times m$ matrices $\mathbf{Z}, \mathbf{G}, \mathbf{H}$ of rank at most $2r$, we have*

$$\left| \frac{2}{\alpha+\beta}[\nabla^2 f(\mathbf{Z})](\mathbf{G},\mathbf{H}) - \langle \mathbf{G},\mathbf{H}\rangle \right| \le \frac{\beta-\alpha}{\beta+\alpha} \|\mathbf{G}\|_F \|\mathbf{H}\|_F .$$

Proof of Proposition 5.4.1. This proof follows similar steps to the proof of [78, Lemma 2.1]. First note that the bilinear form $[\nabla^2 f(\mathbf{Z})](\mathbf{G},\mathbf{H}) = \sum_{i,j,k,l} \frac{\partial^2 f(\mathbf{Z})}{\partial \mathbf{Z}_{ij} \partial \mathbf{Z}_{kl}} \mathbf{G}_{ij}\mathbf{H}_{kl}$ implies $[\nabla^2 f(\mathbf{Z})](\mathbf{G},\mathbf{H})$ is invariant under all scalings for both $\mathbf{G}$ and $\mathbf{H}$, i.e.,

$$[\nabla^2 f(\mathbf{Z})](a\mathbf{G}, b\mathbf{H}) = ab[\nabla^2 f(\mathbf{Z})](\mathbf{G},\mathbf{H})$$

for any $a, b \in \mathbb{R}$. If either $\mathbf{G}$ or $\mathbf{H}$ is zero, (5.3) holds since both sides are 0.

Now suppose both $\mathbf{G}$ or $\mathbf{H}$ are nonzero. By the scaling invariance property of both sides in (5.3), we assume $\|\mathbf{G}\|_F = \|\mathbf{H}\|_F = 1$ without loss of generality. Note that the $(2r, 4r)$-restricted strong convexity and smoothness condition (5.3) implies

$$\alpha \|\mathbf{G} \pm \mathbf{H}\|_F^2 \le [\nabla^2 f(\mathbf{X})](\mathbf{G} \pm \mathbf{H}, \mathbf{G} \pm \mathbf{H}) \le \beta \|\mathbf{G} \pm \mathbf{H}\|_F^2 .$$

Thus we have

$$\begin{aligned} -\frac{\beta-\alpha}{2}\left(\|\mathbf{G}\|_F^2 + \|\mathbf{H}\|_F^2\right) &\le 2\left[\nabla^2 f(\mathbf{Z})\right](\mathbf{G},\mathbf{H}) - (\alpha+\beta)\langle \mathbf{G},\mathbf{H}\rangle \\ &\le \frac{\beta-\alpha}{2}\left(\|\mathbf{G}\|_F^2 + \|\mathbf{H}\|_F^2\right), \end{aligned}$$

which further implies

$$\left|2\left[\nabla^2 f(\mathbf{Z})\right](\mathbf{G},\mathbf{H}) - (\alpha+\beta)\langle \mathbf{G},\mathbf{H}\rangle\right| \le \beta - \alpha = (\beta-\alpha)\|\mathbf{G}\|_F \|\mathbf{H}\|_F .$$

□

D.3 Proof of Lemma 5.4.1

Lemma D.3.1 (Lemma 5.4.1)**.** *Suppose $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (5.3)*. For any critical point $\mathbf{W}$ of* (5.5)*, let $\mathbf{P}_{\mathbf{W}} \in \mathbb{R}^{(m+n)\times(m+n)}$ be the orthogonal projector onto the column space of $\mathbf{W}$. Then*

$$\left\|(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{P}_{\mathbf{W}}\right\|_F \le 2\frac{\beta-\alpha}{\beta+\alpha}\|\mathbf{X} - \mathbf{X}^\star\|_F .$$

Proof of Lemma 5.4.1. First recall the notation $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{X}^\star = \mathbf{U}^\star\mathbf{V}^\star$, and

$$\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}, \widehat{\mathbf{W}} = \begin{bmatrix}\mathbf{U}\\ -\mathbf{V}\end{bmatrix}, \mathbf{W}^\star = \begin{bmatrix}\mathbf{U}^\star\\ \mathbf{V}^\star\end{bmatrix}, \widehat{\mathbf{W}}^\star = \begin{bmatrix}\mathbf{U}^\star\\ -\mathbf{V}^\star\end{bmatrix}.$$

It follows from (5.17) and (5.18) that any critical point $\mathbf{W}$ satisfies

$$\begin{bmatrix}\mathbf{0} & \nabla f(\mathbf{X})\\ \nabla f(\mathbf{X})^\top & \mathbf{0}\end{bmatrix}\mathbf{W} = \mathbf{0},$$

which gives

$$\begin{aligned}
0 &= \langle \begin{bmatrix}\mathbf{0} & \nabla f(\mathbf{X})\\ \nabla f(\mathbf{X})^\top & \mathbf{0}\end{bmatrix}, \mathbf{Z}\mathbf{W}^\top\rangle\\
&= \langle \begin{bmatrix}\mathbf{0} & \nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^\star)\\ \nabla f(\mathbf{X})^\top - \nabla f(\mathbf{X}^\star)^\top & \mathbf{0}\end{bmatrix}, \mathbf{Z}\mathbf{W}^\top\rangle\\
&= \underbrace{\langle \nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^\star) - \frac{\alpha+\beta}{2}(\mathbf{X}-\mathbf{X}^\star), \mathbf{Z}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{Z}_\mathbf{V}^\top\rangle}_{\daleth_1} + \frac{\alpha+\beta}{2}\underbrace{\langle \mathbf{X}-\mathbf{X}^\star, \mathbf{Z}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{Z}_\mathbf{V}^\top\rangle}_{\daleth_2}
\end{aligned} \tag{D.1}$$

for any $\mathbf{Z} = \begin{bmatrix}\mathbf{Z}_\mathbf{U}\\ \mathbf{Z}_\mathbf{V}\end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$. Here the second line utilizes the fact $\nabla f(\mathbf{X}^\star) = \mathbf{0}$. We bound $\daleth_1$ by first using integral form of the mean value theorem for $\nabla f(\mathbf{X})$:

$$\daleth_1 = \int_0^1 \left[\nabla^2 f(t\mathbf{X} + (1-t)\mathbf{X}^\star)\right](\mathbf{X}-\mathbf{X}^\star, \mathbf{Z}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{Z}_\mathbf{V}^\top)dt - \frac{\alpha+\beta}{2}\left\langle \mathbf{X}-\mathbf{X}^\star, \mathbf{Z}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{Z}_\mathbf{V}^\top\right\rangle.$$

Noting that all the three matrices $t\mathbf{X} + (1-t)\mathbf{X}^\star$, $\mathbf{X}-\mathbf{X}^\star$ and $\mathbf{Z}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{Z}_\mathbf{V}^\top$ have rank at most $2r$, it follows from Proposition 5.4.1 that

$$|\daleth_1| \le \frac{\beta-\alpha}{2}\left\|\mathbf{X}-\mathbf{X}^\star\right\|_F \left\|\mathbf{Z}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{Z}_\mathbf{V}^\top\right\|_F,$$

which when plugged into (D.1) gives

$$\begin{aligned}
\frac{\alpha+\beta}{2}\daleth_2 &= -\daleth_1\\
&\le \frac{\beta-\alpha}{2}\left\|\mathbf{X}-\mathbf{X}^\star\right\|_F \left\|\mathbf{Z}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{Z}_\mathbf{V}^\top\right\|_F.
\end{aligned} \tag{D.2}$$

Now let $\mathbf{Z} = (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{W}^{T\dagger}$, which gives $\mathbf{Z}\mathbf{W}^\top = (\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top})\mathbf{P}_\mathbf{W}$. Here $\dagger$ denotes the pseudoinverse of a matrix and $\mathbf{P}_\mathbf{W}$ is the orthogonal projector onto the range of $\mathbf{W}$. Utilizing the fact $\widehat{\mathbf{W}}^\top\mathbf{W} = \mathbf{0}$

from (5.7), we further connect the left hand side of (D.2) with $\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{P}_{\mathbf{W}}\right\|_F^2$ by

$$\begin{aligned}\frac{\alpha+\beta}{2}\urcorner_2 =& \frac{\alpha+\beta}{2}\urcorner_2 + \frac{\alpha+\beta}{4}\langle\widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \mathbf{Z}\mathbf{W}^\top\rangle \\ =& \frac{\alpha+\beta}{4}\left\langle\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}, \left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{P}_{\mathbf{W}}\right\rangle \\ &+ \frac{\alpha+\beta}{4}\left\langle\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{P}_{\mathbf{W}}\right\rangle \\ \geq& \frac{\alpha+\beta}{4}\left\langle\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}, \left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{P}_{\mathbf{W}}\right\rangle \\ =& \frac{\alpha+\beta}{4}\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{P}_{\mathbf{W}}\right\|_F^2,\end{aligned} \tag{D.3}$$

where the inequality follows because

$$\left\langle\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \mathbf{W}^\star\mathbf{W}^{\star\top}\mathbf{P}_{\mathbf{W}}\right\rangle = 0 \quad (\text{since } \widehat{\mathbf{W}}^{\star\top}\widehat{\mathbf{W}}^\star = \mathbf{0})$$

and

$$\left\langle\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \mathbf{W}\mathbf{W}^\top\mathbf{P}_{\mathbf{W}}\right\rangle = \left\langle\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \mathbf{W}\mathbf{W}^\top\right\rangle \geq 0.$$

On the other hand, we give an upper bound on the right hand side of (D.2):

$$\begin{aligned}\|\mathbf{X} - \mathbf{X}^\star\|_F\left\|\mathbf{Z}_{\mathbf{U}}\mathbf{V}^\top + \mathbf{U}\mathbf{Z}_{\mathbf{V}}^\top\right\|_F &\leq \|\mathbf{X} - \mathbf{X}^\star\|_F\sqrt{2\left\|\mathbf{Z}_{\mathbf{U}}\mathbf{V}^\top\right\|_F^2 + 2\left\|\mathbf{U}\mathbf{Z}_{\mathbf{V}}^\top\right\|_F^2} \\ &\leq \|\mathbf{X} - \mathbf{X}^\star\|_F\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{P}_{\mathbf{W}}\right\|_F,\end{aligned}$$

where the last line follows because $\left\|\mathbf{Z}_{\mathbf{U}}\mathbf{V}^\top\right\|_F^2 + \left\|\mathbf{Z}_{\mathbf{V}}\mathbf{U}^\top\right\|_F^2 = \left\|\mathbf{Z}_{\mathbf{U}}\mathbf{U}^\top\right\|_F^2 + \left\|\mathbf{Z}_{\mathbf{V}}\mathbf{V}^\top\right\|_F^2$ (since $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}$), implying $2\left\|\mathbf{Z}_{\mathbf{U}}\mathbf{V}^\top\right\|_F^2 + 2\left\|\mathbf{U}\mathbf{Z}_{\mathbf{V}}^\top\right\|_F^2 = \left\|\mathbf{Z}\mathbf{W}^\top\right\|_F^2$. This together with (D.2) and (D.3) completes the proof. □

D.4 Proof of Lemma 5.4.2

Lemma D.4.1 (Lemma 5.4.2)**.** *For any matrices $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{n\times r}$ with ranks r_1 and r_2, respectively, let $\mathbf{R} = \arg\min_{\mathbf{R}'\in\mathcal{O}_r}\|\mathbf{C} - \mathbf{D}\mathbf{R}'\|_F$. Then*

$$\|\mathbf{C}\mathbf{C}^\top - \mathbf{D}\mathbf{D}^\top\|_F^2/\|\mathbf{C} - \mathbf{D}\mathbf{R}\|_F^2 \geq \max\left\{2(\sqrt{2}-1)\sigma_r^2(\mathbf{D}), \min\left\{\sigma_{r_1}^2(\mathbf{C}), \sigma_{r_2}^2(\mathbf{D})\right\}\right\}.$$

If $\mathbf{C} = \mathbf{0}$, then we have

$$\left\|\mathbf{C}\mathbf{C}^\top - \mathbf{D}\mathbf{D}^\top\right\|_F^2 \geq \sigma_{r_2}^2(\mathbf{D})\|\mathbf{C} - \mathbf{D}\mathbf{R}\|_F^2.$$

Proof of Lemma 5.4.2. When $\mathbf{C} \neq \mathbf{0}$, the proof follows directly from the following results.

Lemma D.4.2. *[6, Lemma 3] For any matrices $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{n\times r}$ with rank r_1 and r_2, respectively, let $\mathbf{R} = \arg\min_{\widetilde{\mathbf{R}}\in\mathcal{O}_r}\|\mathbf{C} - \mathbf{D}\mathbf{R}\|_F$. Then*

$$\left\|\mathbf{C}\mathbf{C}^\top - \mathbf{D}\mathbf{D}^\top\right\|_F \geq \min\left\{\sigma_{r_1}(\mathbf{C}), \sigma_{r_2}(\mathbf{D})\right\} \cdot \|\mathbf{C} - \mathbf{D}\mathbf{R}\|_F .$$

Lemma D.4.3. *[102, Lemma 5.4] For any matrices $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{n\times r}$ with $\operatorname{rank}(\mathbf{D}) = r$, let $\mathbf{R} = \arg\min_{\widetilde{\mathbf{R}}\in\mathcal{O}_r} \|\mathbf{C} - \mathbf{D}\mathbf{R}\|_F$. Then*

$$\left\|\mathbf{C}\mathbf{C}^\top - \mathbf{D}\mathbf{D}^\top\right\|_F^2 \geq 2(\sqrt{2}-1)\sigma_r^2(\mathbf{D}) \|\mathbf{C} - \mathbf{D}\mathbf{R}\|_F^2 .$$

If $\mathbf{C} = \mathbf{0}$, then we have

$$\left\|\mathbf{C}\mathbf{C}^\top - \mathbf{D}\mathbf{D}^\top\right\|_F^2 = \left\|\mathbf{D}\mathbf{D}^\top\right\|_F^2 = \sum_{i=1}^{r_2} \sigma_i^4(\mathbf{D}) \geq \sigma_{r_2}^2(\mathbf{D}) \sum_{i=1}^{r_2} \sigma_i^2(\mathbf{D}) = \sigma_{r_2}^2(\mathbf{D}) \|\mathbf{C} - \mathbf{D}\mathbf{R}\|_F^2 .$$

□

D.5 Proof of Eq. (5.20)

Proof of Eq. (5.20). We prove the upper bounds for the four terms as follows.

Bounding term Π_1. Utilizing the fact that $\boldsymbol{\Delta}_\mathbf{U} = \mathbf{U} - \mathbf{U}^\star\mathbf{R}$ and $\boldsymbol{\Delta}_\mathbf{V} = \mathbf{V} - \mathbf{V}^\star\mathbf{R}$, we have

$$\begin{aligned}
\Pi_1 &= \left\langle \nabla f(\mathbf{X}), \boldsymbol{\Delta}_\mathbf{U}\boldsymbol{\Delta}_\mathbf{V}^\top \right\rangle \\
&= \left\langle \nabla f(\mathbf{X}), (\mathbf{U} - \mathbf{U}^\star\mathbf{R})(\mathbf{V} - \mathbf{V}^\star\mathbf{R})^\top \right\rangle \\
&= \left\langle \nabla f(\mathbf{X}), \mathbf{X} + \mathbf{X}^\star - \mathbf{U}^\star\mathbf{R}^\top\mathbf{V}^T - \mathbf{U}\mathbf{R}^\top\mathbf{V}^{\star\top} \right\rangle \\
&\overset{(i)}{=} -\left\langle \nabla f(\mathbf{X}), \mathbf{X} - \mathbf{X}^\star \right\rangle \\
&\overset{(ii)}{=} -\left\langle \nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^\star), \mathbf{X} - \mathbf{X}^\star \right\rangle \\
&\overset{(iii)}{\leq} -\alpha \|\mathbf{X} - \mathbf{X}^\star\|_F^2 ,
\end{aligned}$$

where (i) follows from (5.17) and (5.18), (ii) utilizes $\nabla f(\mathbf{X}^\star) = \mathbf{0}$, and (iii) follows by using the $(2r, 4r)$-restricted strict convexity property (5.3):

$$\begin{aligned}
\left\langle \nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^\star), \mathbf{X} - \mathbf{X}^\star \right\rangle &= \int_0^1 \left[\nabla^2 f(t\mathbf{X} + (1-t)\mathbf{X}^\star)\right](\mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star)\, dt \\
&\geq \int_0^1 \alpha \left\langle \mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star \right\rangle dt \\
&= \alpha \|\mathbf{X} - \mathbf{X}^\star\|_F^2 ,
\end{aligned}$$

where the first line follows from the integral form of the mean value theorem for vector-valued functions, and the second line uses the fact that both $t\mathbf{X} + (1-t)\mathbf{X}^\star$ and $\mathbf{X} - \mathbf{X}^\star$ have rank at most $2r$, and the $(2r, 4r)$-restricted strong convexity of the Hessian $\nabla^2 f(\cdot)$.

Bounding term Π_2. By the smoothness condition (5.3), we have

$$\begin{aligned}\Pi_2 &= \left[\nabla^2 f(\mathbf{X})\right]\left(\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top + \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top, \boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top + \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right)\\ &\le \beta\left\|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top + \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\|_F^2\\ &\le 2\beta\left(\left\|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top\right\|_F^2 + \left\|\mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\|_F^2\right)\\ &= \beta\left\|\mathbf{W}\boldsymbol{\Delta}^\top\right\|_F^2,\end{aligned}$$

where the last line holds because $\left\|\mathbf{D}\mathbf{U}^\top\right\|_F^2 = \left\|\mathbf{D}\mathbf{V}^\top\right\|_F^2$ for any $\mathbf{D} \in \mathbb{R}^{p\times r}$ with arbitrary $p \ge 1$ since any critical point $\mathbf{W}$ satisfies $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}$.

Bounding term Π_3.

$$\begin{aligned}\Pi_3 &= \langle \mathbf{U}\boldsymbol{\Delta}_{\mathbf{U}}^\top, \boldsymbol{\Delta}_{\mathbf{U}}\mathbf{U}^\top\rangle + \langle \mathbf{V}\boldsymbol{\Delta}_{\mathbf{V}}^\top, \boldsymbol{\Delta}_{\mathbf{V}}\mathbf{V}^\top\rangle - 2\langle \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top, \boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top\rangle\\ &\le \left\|\mathbf{U}\boldsymbol{\Delta}_{\mathbf{U}}^\top\right\|_F^2 + \left\|\mathbf{V}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\|_F^2 + \left\|\mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\|_F^2 + \left\|\mathbf{V}\boldsymbol{\Delta}_{\mathbf{U}}^\top\right\|_F^2\\ &= \left\|\mathbf{W}\boldsymbol{\Delta}^\top\right\|_F^2.\end{aligned}$$

Bounding term Π_4.

$$\begin{aligned}\Pi_4 &= \left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, (\mathbf{W} - \mathbf{W}^\star\mathbf{R})(\mathbf{W} - \mathbf{W}^\star\mathbf{R})^\top\right\rangle\\ &\overset{(i)}{=} -\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\rangle\\ &\overset{(ii)}{\le} -\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\rangle + \left\langle \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\rangle\\ &= -\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top - \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\rangle \le 2\left\|\mathbf{X} - \mathbf{X}^\star\right\|_F^2,\end{aligned}$$

where (i) holds because $\widehat{\mathbf{W}}^\top\mathbf{W} = \mathbf{0}$, and (ii) follows because $\widehat{\mathbf{W}}^{\star\top}\mathbf{W}^\star = \mathbf{0}$ and $\langle \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \mathbf{W}\mathbf{W}^\top\rangle \ge 0$. □

D.6 Proof of Eq. (5.22)

Proof of Eq. (5.22). To show (5.22), expanding the left hand side of (5.22), it is equivalent to show

$$\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top}\right\|_F^2 + \left\|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\right\|_F^2 \le 2\left\|\mathbf{X} - \mathbf{X}^\star\right\|_F^2.$$

Expanding both sides of the above equation and utilizing the fact $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}$ and $\mathbf{U}^{\star\top}\mathbf{U}^\star = \mathbf{V}^{\star\top}\mathbf{V}^\star$, the remaining step is to show

$$\operatorname{tr}\left(\mathbf{U}\mathbf{U}^\top\mathbf{U}^\star\mathbf{U}^{\star\top}\right) + \left(\mathbf{V}\mathbf{V}^\top\mathbf{V}^\star\mathbf{V}^{\star\top}\right) \ge 2\operatorname{tr}\left(\mathbf{U}\mathbf{V}^\top\mathbf{V}^\star\mathbf{U}^{\star\top}\right).$$

Thus, we obtain (5.22) by noting that the above equation is equivalent to $\operatorname{tr}\left(\left(\mathbf{U}^{\star\top}\mathbf{U} - \mathbf{V}^{\star\top}\mathbf{V}\right)^2\right) \ge 0$. □

APPENDIX E

APPENDICES FOR CHAPTER 6

E.1 The optimization geometry of low-rank matrix factorization

In this appendix, we consider the low-rank matrix factorization problem

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}}\; g(\mathbf{W}) := \frac{1}{2}\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\right\|_F^2 + \rho(\mathbf{W}) \tag{E.1}$$

where $\rho(\mathbf{W})$ is the regularizer used in (6.9) and repeated here:

$$\rho(\mathbf{W}) = \frac{\mu}{4}\left\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right\|_F^2.$$

We provide a comprehensive geometric analysis for the matrix factorization problem (E.1). In particular, we show that the objective function in (E.1) obeys the strict saddle property and has no spurious local minima not only for exact-parameterization ($r = \mathrm{rank}(\mathbf{X}^\star)$), but also for over-parameterization ($r > \mathrm{rank}(\mathbf{X}^\star)$) and under-parameterization ($r < \mathrm{rank}(\mathbf{X}^\star)$). For the exact-parameterization case, we further show that the objective function satisfies the robust strict saddle property, ensuring global convergence of many local search algorithms in polynomial time. As we believe these results are also of independent interest and to make it easy to follow, we only present the main results in this appendix and defer the proofs to other appendices.

E.1.1 Relationship to PSD low-rank matrix factorization

Similar to (6.8), let $\mathbf{X}^\star = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Psi}^\top = \sum_{i=1}^r \sigma_i\boldsymbol{\phi}_i\boldsymbol{\psi}_i^\top$ be a reduced SVD of $\mathbf{X}^\star$, where $\mathbf{\Sigma}$ is a diagonal matrix with $\sigma_1 \geq \cdots \geq \sigma_r$ along its diagonal, and denote $\mathbf{U}^\star = \mathbf{\Phi}\mathbf{\Sigma}^{1/2}\mathbf{R}$, $\mathbf{V}^\star = \mathbf{\Psi}\mathbf{\Sigma}^{1/2}\mathbf{R}$ for any $\mathbf{R} \in \mathcal{O}_r$. The following result to some degree characterizes the relationship between the nonsymmetric low-rank matrix factorization problem (E.1) and the following PSD low-rank matrix factorization problem [104]:

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\text{minimize}}\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{M}\right\|_F^2 \tag{E.2}$$

where $\mathbf{M} \in \mathbb{R}^{n\times n}$ is a rank-r PSD matrix.

Lemma E.1.1. *Suppose $g(\mathbf{W})$ is defined as in* (E.1) *with $\mu > 0$. Then we have*

$$g(\mathbf{W}) \geq \min\{\frac{\mu}{4}, \frac{1}{8}\}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2.$$

In particular, if we choose $\mu = \frac{1}{2}$, then we have

$$g(\mathbf{W}) = \frac{1}{8}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2 + \frac{1}{4}\left\|\mathbf{U}^\top\mathbf{U}^\star - \mathbf{V}^\top\mathbf{V}^\star\right\|_F^2.$$

The proof of Lemma E.1.1 is given in Appendix E.5. Informally, Lemma E.1.1 indicates that minimizing $g(\mathbf{W})$ also results in minimizing $\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2$ (which is the same form as the objective function in (E.2)) and hence the distance between $\mathbf{W}$ and $\mathbf{W}^\star$ (though $\mathbf{W}^\star$ is unavailable at priori). The global geometry for the PSD low-rank matrix factorization problem (E.2) is recently analyzed by Li et al. in [104].

E.1.2 Characterization of critical points

We first provide the gradient and Hessian expression for $g(\mathbf{W})$. The gradient of $g(\mathbf{W})$ is given by

$$\begin{aligned}\nabla_{\mathbf{U}} g(\mathbf{U},\mathbf{V}) &= (\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star)\mathbf{V} + \mu\mathbf{U}(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}),\\ \nabla_{\mathbf{V}} g(\mathbf{U},\mathbf{V}) &= (\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star)^\top\mathbf{U} - \mu\mathbf{V}(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}),\end{aligned}$$

which can be rewritten as

$$\nabla g(\mathbf{W}) = \begin{bmatrix}(\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star)\mathbf{V}\\ (\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star)^\top\mathbf{U}\end{bmatrix} + \mu\widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top\mathbf{W}.$$

Standard computations give the Hessian quadrature form $[\nabla^2 g(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta})$ for any $\mathbf{\Delta} = \begin{bmatrix}\mathbf{\Delta}_{\mathbf{U}}\\ \mathbf{\Delta}_{\mathbf{V}}\end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$ (where $\mathbf{\Delta}_{\mathbf{U}} \in \mathbb{R}^{n\times r}$ and $\mathbf{\Delta}_{\mathbf{V}} \in \mathbb{R}^{m\times r}$) as

$$[\nabla^2 g(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta}) = \left\|\mathbf{\Delta}_{\mathbf{U}}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_{\mathbf{V}}^\top\right\|_F^2 + 2\left\langle \mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star, \mathbf{\Delta}_{\mathbf{U}}\mathbf{\Delta}_{\mathbf{V}}^\top\right\rangle + [\nabla^2\rho(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta}), \tag{E.3}$$

where

$$[\nabla^2\rho(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta}) = \mu\left\langle \widehat{\mathbf{W}}^\top\mathbf{W}, \widehat{\mathbf{\Delta}}^\top\mathbf{\Delta}\right\rangle + \mu\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{\Delta}}^\top, \mathbf{\Delta}\mathbf{W}^\top\right\rangle + \mu\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \mathbf{\Delta}\mathbf{\Delta}^\top\right\rangle. \tag{E.4}$$

By Lemma 6.3.1, we can simplify the equations for critical points as follows

$$\nabla_{\mathbf{U}}\rho(\mathbf{U},\mathbf{V}) = \mathbf{U}\mathbf{U}^\top\mathbf{U} - \mathbf{X}^\star\mathbf{V} = \mathbf{0}, \tag{E.5}$$
$$\nabla_{\mathbf{V}}\rho(\mathbf{U},\mathbf{V}) = \mathbf{V}\mathbf{V}^\top\mathbf{V} - \mathbf{X}^{\star\top}\mathbf{U} = \mathbf{0}. \tag{E.6}$$

Now suppose $\mathbf{W}$ is a critical point of $g(\mathbf{W})$. We can apply the Gram-Schmidt process to orthonormalize the columns of $\mathbf{U}$ such that $\widetilde{\mathbf{U}} = \mathbf{U}\mathbf{R}$, where $\widetilde{\mathbf{U}}$ is orthogonal and $\mathbf{R} \in \mathcal{O}_r = \left\{\mathbf{R} \in \mathbb{R}^{r\times r}, \mathbf{R}^\top\mathbf{R} = \mathbf{I}\right\}$.[44] Also let $\widetilde{\mathbf{V}} = \mathbf{V}\mathbf{R}$. Since $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}$, we have $\widetilde{\mathbf{U}}^\top\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}}^\top\widetilde{\mathbf{V}}$. Thus $\widetilde{\mathbf{V}}$ is also orthogonal. Noting that $\mathbf{U}\mathbf{V}^\top = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top$, we conclude that $g(\mathbf{W}) = g(\widetilde{\mathbf{W}})$ and $\widetilde{\mathbf{W}}$ is also a critical point of $g(\mathbf{W})$ since $\nabla_{\widetilde{\mathbf{U}}} g(\widetilde{\mathbf{W}}) = \nabla_{\mathbf{U}} g(\mathbf{W})\mathbf{R} = \mathbf{0}$ and $\nabla_{\widetilde{\mathbf{V}}} g(\widetilde{\mathbf{W}}) = \nabla_{\mathbf{V}} g(\mathbf{W})\mathbf{R} = \mathbf{0}$. Also for any $\mathbf{\Delta} \in \mathbb{R}^{(n+m)\times r}$, we have $[\nabla^2 g(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta}) = [\nabla^2 g(\widetilde{\mathbf{W}})](\mathbf{\Delta}\mathbf{R},\mathbf{\Delta}\mathbf{R})$,

[44] Another way to find $\mathbf{R}$ is via the SVD. Let $\mathbf{U} = \mathbf{L}\mathbf{\Sigma}\mathbf{R}^\top$ be a reduced SVD of $\mathbf{U}$, where $\mathbf{L}$ is an $n \times r$ orthonormal matrix, $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix with non-negative diagonals, and $\mathbf{R} \in \mathcal{O}_r$. Then $\widetilde{\mathbf{U}} = \mathbf{U}\mathbf{R} = \mathbf{L}\mathbf{\Sigma}$ is orthogonal, but has possible zero columns.

indicating that $\mathbf{W}$ and $\widetilde{\mathbf{W}}$ have the same Hessian information. Thus, without loss of generality, we assume $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, but possibly include zero columns. With this, we use $\mathbf{u}_i$ and $\mathbf{v}_i$ to denote the i-th columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. It follows from $\nabla g(\mathbf{W}) = \mathbf{0}$ that

$$\begin{aligned}\|\mathbf{u}_i\|^2\mathbf{u}_i &= \mathbf{X}^\star\mathbf{v}_i,\\ \|\mathbf{v}_i\|^2\mathbf{v}_i &= \mathbf{X}^{\star\top}\mathbf{u}_i,\end{aligned}$$

which indicates that

$$(\mathbf{u}_i, \mathbf{v}_i) \in \left\{(\sqrt{\lambda_1}\mathbf{p}_1, \sqrt{\lambda_1}\mathbf{q}_1), \ldots, (\sqrt{\lambda_r}\mathbf{p}_r, \sqrt{\lambda_r}\mathbf{q}_r), (\mathbf{0}, \mathbf{0})\right\}.$$

Now we identify all the critical points of $g(\mathbf{W})$ in the following lemma, which is formally proved with an algebraic approach in Appendix E.6.

Lemma E.1.2. *Let $\mathbf{X}^\star = \mathbf{\Phi\Sigma\Psi}^\top = \sum_{i=1}^r \sigma_i\boldsymbol{\phi}_i\boldsymbol{\psi}_i^\top$ be a reduced SVD of $\mathbf{X}^\star$ and $g(\mathbf{W})$ be defined as in* (E.1) *with $\mu > 0$. Any $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ is a critical point of $g(\mathbf{W})$ if and only if $\mathbf{W} \in \mathcal{C}$ with*

$$\mathcal{C} := \left\{\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi\Lambda}^{1/2}\mathbf{R}, \mathbf{R} \in \mathcal{O}_r, \mathbf{\Lambda} \text{ is diagonal}, \mathbf{\Lambda} \geq \mathbf{0}, (\mathbf{\Sigma} - \mathbf{\Lambda})\mathbf{\Sigma} = \mathbf{0}\right\}. \tag{E.7}$$

Intuitively, (E.7) means that a critical point $\mathbf{W}$ of $g(\mathbf{W})$ is one such that $\mathbf{UV}^\top$ is a rank-ℓ approximation to $\mathbf{X}^\star$ with $\ell \leq r$ and $\mathbf{U}$ and $\mathbf{V}$ are equal factors of this rank-ℓ approximation. Let $\lambda_1, \lambda_2, \ldots, \lambda_r$ denote the diagonals of $\mathbf{\Lambda}$. Unlike $\mathbf{\Sigma}$, we note that these diagonals $\lambda_1, \lambda_2, \ldots, \lambda_r$ are not necessarily placed in decreasing or increasing order. Actually, this equation $(\mathbf{\Sigma} - \mathbf{\Lambda})\mathbf{\Sigma} = \mathbf{0}$ is equivalent to

$$\lambda_i \in \{\sigma_i, 0\}$$

for all $i \in \{1, 2, \ldots, r\}$. Further, we introduce the set of optimal solutions:

$$\mathcal{X} := \left\{\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi\Sigma}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi\Sigma}^{1/2}\mathbf{R}, \mathbf{R} \in \mathcal{O}_r\right\}. \tag{E.8}$$

It is clear that the set $\mathcal{X}$ containing all the optimal solutions, the set $\mathcal{C}$ containing all the critical points and the set $\mathcal{E}$ containing all the points with balanced factors have the nesting relationship: $\mathcal{X} \subset \mathcal{C} \subset \mathcal{E}$. Before moving to the next section, we provide one more result regarding $\mathbf{W} \in \mathcal{E}$. The proof of the following result is given in Appendix E.7.

Lemma E.1.3. *For any $\mathbf{\Delta} = \begin{bmatrix}\mathbf{\Delta}_\mathbf{U}\\ \mathbf{\Delta}_\mathbf{V}\end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$ and $\mathbf{W} \in \mathcal{E}$ where $\mathcal{E}$ is defined in* (6.10)*, we have*

$$\|\mathbf{\Delta}_{\mathbf{U}}\mathbf{U}^\top\|_F^2 + \|\mathbf{\Delta}_{\mathbf{V}}\mathbf{V}^\top\|_F^2 = \|\mathbf{\Delta}_{\mathbf{U}}\mathbf{V}^\top\|_F^2 + \|\mathbf{\Delta}_{\mathbf{V}}\mathbf{U}^\top\|_F^2, \tag{E.9}$$

and

$$\nabla^2\rho(\mathbf{W}) \succeq \mathbf{0}. \tag{E.10}$$

E.1.3 Strict saddle property

Lemma E.1.3 implies that the Hessian of $\rho(\mathbf{W})$ evaluated at any critical point $\mathbf{W}$ is PSD, i.e., $\nabla^2\rho(\mathbf{W}) \succeq \mathbf{0}$ for all $\mathbf{W} \in \mathcal{C}$. Despite this fact, the following result establishes the strict saddle property for $g(\mathbf{W})$.

Theorem E.1.1. *Let $g(\mathbf{W})$ be defined as in* (E.1) *with $\mu > 0$ and $\mathrm{rank}(\mathbf{X}^\star) = r$. Let $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ be any critical point satisfying $\nabla g(\mathbf{W}) = \mathbf{0}$, i.e., $\mathbf{W} \in \mathcal{C}$. Any $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$ is a strict saddle of $g(\mathbf{W})$ satisfying*

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -\frac{1}{2}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\| \leq -\sigma_r(\mathbf{X}^\star). \tag{E.11}$$

Furthermore, $g(\mathbf{W})$ is not strongly convex at any global minimum point $\mathbf{W} \in \mathcal{X}$.

The proof of Theorem E.1.1 is given in Appendix E.8. We note that this strict saddle property is also covered in [8, Theorem 3], but with much looser bounds (in particular, directly applying [8, Theorem 3] gives $\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -0.1\sigma_r(\mathbf{X}^\star)$ rather than $\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -\sigma_r(\mathbf{X}^\star)$ in (E.11)). Theorem E.1.1 actually implies that $g(\mathbf{W})$ has no spurious local minima (since all local minima belong to $\mathcal{X}$) and obeys the strict saddle property. With the strict saddle property and lack of spurious local minima for $g(\mathbf{W})$, the recent results [178, 179] ensure that gradient descent converges to a global minimizer almost surely with random initialization. We also note that Theorem E.1.1 states that $g(\mathbf{W})$ is not strongly convex at any global minimum point $\mathbf{W} \in \mathcal{X}$ because of the invariance property of $g(\mathbf{W})$. This is the reason we introduce the distance in (6.12) and also the robust strict saddle property in Definition 6.2.9.

E.1.4 Extension to over-parameterized case: $\mathrm{rank}(\mathbf{X}^\star) < r$

In this section, we briefly discuss the over-parameterized scenario where the low-rank matrix $\mathbf{X}^\star$ has rank smaller than r. Similar to Theorem E.1.1, the following result shows that the strict saddle property also holds in this case.

Theorem E.1.2. *Let $\mathbf{X}^\star = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Psi}^\top = \sum_{i=1}^{r'}\sigma_i\boldsymbol{\phi}_i\boldsymbol{\psi}_i^\top$ be a reduced SVD of $\mathbf{X}^\star$ with $r' \leq r$, and let $g(\mathbf{W})$ be defined as in* (E.1) *with $\mu > 0$. Any $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ is a critical point of $g(\mathbf{W})$ if and only if $\mathbf{W} \in \mathcal{C}$ with*

$$\mathcal{C} := \left\{\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi}\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{R}\mathbf{R}^\top = \mathbf{I}_{r'}, \mathbf{\Lambda} \text{ is diagonal}, \mathbf{\Lambda} \geq \mathbf{0}, (\mathbf{\Sigma} - \mathbf{\Lambda})\mathbf{\Sigma} = \mathbf{0}\right\}$$

Further, all the local minima (which are also global) belong to the following set

$$\mathcal{X} = \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \mathbf{\Phi}\mathbf{\Sigma}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}\mathbf{\Sigma}^{1/2}\mathbf{R}, \mathbf{R}\mathbf{R}^\top = \mathbf{I}_{r'} \right\}$$

Finally, any $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$ *is a strict saddle of* $g(\mathbf{W})$ *satisfying*

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -\frac{1}{2} \left\| \mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star \mathbf{W}^{\star\top} \right\| \leq -\sigma_{r'}(\mathbf{X}^\star).$$

The proof of Theorem E.1.2 is given in Appendix E.9. We note that this strict saddle property is also covered in [8, Theorem 3], but with much looser bounds (in particular, directly applying [8, Theorem 3] gives $\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -0.1\sigma_{r'}(\mathbf{X}^\star)$ rather than $\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -\sigma_{r'}(\mathbf{X}^\star)$ in Theorem E.1.2).

E.1.5 Extension to under-parameterized case: $\mathrm{rank}(\mathbf{X}^\star) > r$

We further discuss the under-parameterized case where $\mathrm{rank}(\mathbf{X}^\star) > r$. In this case, (6.3) is also known as the low-rank approximation problem as the product $\mathbf{U}\mathbf{V}^\top$ forms a rank-r approximation to $\mathbf{X}^\star$. Similar to Theorem E.1.1, the following result shows that the strict saddle property also holds for $g(\mathbf{W})$ in this scenario.

Theorem E.1.3. *Let* $\mathbf{X}^\star = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Psi}^\top = \sum_{i=1}^{r'} \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^\top$ *be a reduced SVD of* $\mathbf{X}^\star$ *with* $r' > r$ *and* $\sigma_r(\mathbf{X}^\star) > \sigma_{r+1}(\mathbf{X}^\star)$.[45] *Also let* $g(\mathbf{W})$ *be defined as in* (E.1) *with* $\mu > 0$. *Any* $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ *is a critical point of* $g(\mathbf{W})$ *if and only if* $\mathbf{W} \in \mathcal{C}$ *with*

$$\begin{aligned} \mathcal{C} := \Bigg\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \mathbf{\Phi}[:, \Omega]\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}[:, \Omega]\mathbf{\Lambda}^{1/2}\mathbf{R}, \\ \mathbf{\Lambda} = \mathbf{\Sigma}[\Omega, \Omega], \mathbf{R}\mathbf{R}^\top = \mathbf{I}_\ell, \Omega \subset \{1, 2, \ldots, r'\}, |\Omega| = \ell \leq r \Bigg\} \end{aligned}$$

where we recall that $\mathbf{\Phi}[:, \Omega]$ *is a submatrix of* $\mathbf{\Phi}$ *obtained by keeping the columns indexed by* Ω *and* $\mathbf{\Sigma}[\Omega, \Omega]$ *is an* $\ell \times \ell$ *matrix obtained by taking the elements of* $\mathbf{\Sigma}$ *in rows and columns indexed by* Ω.

Further, all local minima belong to the following set

$$\mathcal{X} = \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{\Lambda} = \mathbf{\Sigma}[1:r, 1:r], \mathbf{R} \in \mathcal{O}_r, \mathbf{U} = \mathbf{\Phi}[:, 1:r]\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}[:, 1:r]\mathbf{\Lambda}^{1/2}\mathbf{R} \right\}.$$

Finally, any $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$ *is a strict saddle of* $g(\mathbf{W})$ *satisfying*

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -(\sigma_r(\mathbf{X}^\star) - \sigma_{r+1}(\mathbf{X}^\star)).$$

[45] If $\sigma_{r_1} = \cdots = \sigma_r = \cdots = \sigma_{r_2}$ with $r_1 \leq r \leq r_2$, then the optimal rank-r approximation to $\mathbf{X}^\star$ is not unique. For this case, the optimal solution set $\mathcal{X}$ for the factorized problem needs to be changed correspondingly, but the main arguments still hold.

The proof of Theorem E.1.3 is given in Appendix E.10. It follows from Eckart-Young-Mirsky theorem [271] that for any $\mathbf{W} \in \mathcal{X}$, $\mathbf{U}\mathbf{V}^\top$ is the best rank-r approximation to $\mathbf{X}^\star$. Thus, this strict saddle property ensures that the local search algorithms applied to the factored problem (E.1) converge to global optimum which corresponds to the best rank-r approximation to $\mathbf{X}^\star$.

E.1.6 Robust strict saddle property

We now consider the revised robust strict saddle property defined in Definition 6.2.9 for the low-rank matrix factorization problem (E.1). As guaranteed by Theorem E.1.1, $g(\mathbf{W})$ satisfies the strict saddle property for any $\mu > 0$. However, too small a μ would make analyzing the robust strict saddle property difficult. To see this, we denote

$$f(\mathbf{W}) = \frac{1}{2}\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\right\|_F^2$$

for convenience. Thus we can rewrite $g(\mathbf{W})$ as the sum of $f(\mathbf{W})$ and $\rho(\mathbf{W})$. Note that for any $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} \in \mathcal{C}$ where $\mathcal{C}$ is the set of critical points defined in (E.7), $\widetilde{\mathbf{W}} = \begin{bmatrix}\mathbf{U}\mathbf{M}\\ \mathbf{V}\mathbf{M}^{-1}\end{bmatrix}$ is a critical point of $f(\mathbf{W})$ for any invertible $\mathbf{M} \in \mathbb{R}^{r\times r}$. This further implies that the gradient at $\widetilde{\mathbf{W}}$ reduces to

$$\nabla g(\widetilde{\mathbf{W}}) = \nabla \rho(\widetilde{\mathbf{W}}),$$

which could be very small if μ is very small since $\rho(\mathbf{W}) = \frac{\mu}{4}\left\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right\|_F^2$. On the other hand, $\widetilde{\mathbf{W}}$ could be far away from any point in $\mathcal{X}$ for some $\mathbf{M}$ that is not well-conditioned. Therefore, we choose a proper μ controlling the importance of the regularization term such that for any $\mathbf{W}$ that is not close to the critical points $\mathcal{X}$, $g(\mathbf{W})$ has large gradient. Motivated by Lemma E.1.1, we choose $\mu = \frac{1}{2}$.

The following result establishes the robust strict saddle property for $g(\mathbf{W})$.

Theorem E.1.4 (Theorem E.11.1)**.** *Let $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3', \mathcal{R}_3'', \mathcal{R}_3'''$ be the regions as defined in Theorem 6.3.1. Let $g(\mathbf{W})$ be defined as in* (E.1) *with $\mu = \frac{1}{2}$. Then $g(\mathbf{W})$ has the following robust strict saddle property:*

1. *For any $\mathbf{W} \in \mathcal{R}_1$, $g(\mathbf{W})$ satisfies local regularity condition:*

$$\langle \nabla g(\mathbf{W}), \mathbf{W} - \mathbf{W}^\star\mathbf{R}\rangle \geq \frac{1}{32}\sigma_r(\mathbf{X}^\star)\,\mathrm{dist}^2(\mathbf{W}, \mathbf{W}^\star) + \frac{1}{48\|\mathbf{X}^\star\|}\left\|\nabla g(\mathbf{W})\right\|_F^2, \tag{E.12}$$

   where $\mathrm{dist}(\mathbf{W}, \mathbf{W}^\star)$ *and* $\mathbf{R}$ *are defined in* (6.12) *and* (6.13)*, respectively.*

2. *For any $\mathbf{W} \in \mathcal{R}_2$, $g(\mathbf{W})$ has a directional negative curvature:*

$$\lambda_{\min}\left(\nabla^2 g(\mathbf{W})\right) \leq -\frac{1}{4}\sigma_r(\mathbf{X}^\star). \tag{E.13}$$

3. For any $\mathbf{W} \in \mathcal{R}_3 = \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$, $g(\mathbf{W})$ has large gradient descent:

$$\|\nabla g(\mathbf{W})\|_F \geq \frac{1}{10}\sigma_r^{3/2}(\mathbf{X}^\star), \quad \forall\, \mathbf{W} \in \mathcal{R}_3'; \tag{E.14}$$

$$\|\nabla g(\mathbf{W})\|_F > \frac{39}{800}\|\mathbf{W}\|^3, \quad \forall\, \mathbf{W} \in \mathcal{R}_3''; \tag{E.15}$$

$$\langle \nabla g(\mathbf{W}), \mathbf{W}\rangle > \frac{1}{20}\left\|\mathbf{W}\mathbf{W}^\top\right\|_F^2, \quad \forall\, \mathbf{W} \in \mathcal{R}_3'''. \tag{E.16}$$

The proof is given in Appendix E.11.

Remark E.1.1. Recall that all the strict saddles of $g(\mathbf{W})$ are actually rank deficient (see Theorem E.1.1). Thus the region $\mathcal{R}_2$ attempts to characterize all the neighbors of the saddle saddles by including all rank deficient points. Actually, (E.13) holds not only for $\mathbf{W} \in \mathcal{R}_2$, but for all $\mathbf{W}$ such that $\sigma_r(\mathbf{W}) \leq \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\mathbf{X}^\star)$. The reason we add another constraint controlling the term $\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F$ is to ensure this negative curvature property in the region $\mathcal{R}_2$ also holds for the matrix sensing problem discussed in next section. This is the same reason we add two more constraints $\|\mathbf{W}\| \leq \frac{20}{19}\|\mathbf{W}^\star\|_F$ and $\|\mathbf{W}\mathbf{W}^\top\|_F \leq \frac{10}{9}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F$ for the region $\mathcal{R}_3'$.

E.2 Proof of Lemma 6.2.1

Lemma E.2.1 (Lemma 6.2.1)**.** *[102, 180] If the function $h(\mathbf{x})$ restricted to a δ neighborhood of $\mathbf{x}^\star$ satisfies the (α, β, δ)-regularity condition, then as long as gradient descent starts from a point $\mathbf{x}_0 \in B(\delta, \mathbf{x}^\star)$, the gradient descent update*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \nu\nabla h(\mathbf{x}_t)$$

with step size $0 < \nu \leq 2\beta$ obeys $\mathbf{x}_t \in B(\delta, \mathbf{x}^\star)$ and

$$\operatorname{dist}^2(\mathbf{x}_t, \mathbf{x}^\star) \leq (1 - 2\nu\alpha)^t \operatorname{dist}^2(\mathbf{x}_0, \mathbf{x}^\star)$$

for all $t \geq 0$.

Proof. Denote $a_{\mathbf{x},\mathbf{x}^\star} = \arg\min_{a' \in \mathcal{G}} \|\mathbf{x} - a'(\mathbf{x}^\star)\|$. Utilizing the definition of distance in (6.4), the regularity condition (6.5) and the assumption that $\mu \leq 2\beta$, we have

$$\begin{aligned}
\operatorname{dist}^2(\mathbf{x}_{t+1}, \mathbf{x}^\star) &= \left\|\mathbf{x}_{t+1} - a_{\mathbf{x}_{t+1},\mathbf{x}^\star}(\mathbf{x}^\star)\right\|^2 \\
&\leq \left\|\mathbf{x}_t - \nu\nabla h(\mathbf{x}_t) - a_{\mathbf{x}_t,\mathbf{x}^\star}(\mathbf{x}^\star)\right\|^2 \\
&= \left\|\mathbf{x}_t - a_{\mathbf{x}_t,\mathbf{x}^\star}(\mathbf{x}^\star)\right\|^2 + \nu^2\left\|\nabla h(\mathbf{x}_t)\right\|^2 - 2\nu\left\langle \mathbf{x}_t - a_{\mathbf{x}_t,\mathbf{x}^\star}(\mathbf{x}^\star), \nabla h(\mathbf{x}_t)\right\rangle \\
&\leq (1 - 2\nu\alpha)\operatorname{dist}^2(\mathbf{x}_t, \mathbf{x}^\star) - \nu(2\beta - \nu)\left\|\nabla h(\mathbf{x}_t)\right\|^2 \\
&\leq (1 - 2\nu\alpha)\operatorname{dist}^2(\mathbf{x}_t, \mathbf{x}^\star)
\end{aligned}$$

where the fourth line uses the regularity condition (6.5) and the last line holds because $\nu \leq 2\beta$. Thus we conclude $\mathbf{x}_t \in B(\delta)$ for all $t \in \mathbb{N}$ if $\mathbf{x}_0 \in B(\delta)$ by noting that $0 \leq 1 - 2\nu\alpha < 1$ since $\alpha\beta \leq \frac{1}{4}$ and $\nu \leq 2\beta$.

□

E.3 Proof of Proposition 6.3.1

Proposition E.3.1 (Proposition 6.3.1)**.** *Suppose $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (6.6) *with positive a and b. Assume $\mathbf{X}^\star$ is a critical point of $f(\mathbf{X})$ with $\mathrm{rank}(\mathbf{X}^\star) = r$. Then $\mathbf{X}^\star$ is the global minimum of* (6.1)*, i.e.,*

$$f(\mathbf{X}^\star) \leq f(\mathbf{X}),\ \forall\, \mathbf{X} \in \mathbb{R}^{n\times m}, \mathrm{rank}(\mathbf{X}) \leq r$$

and the equality holds only at $\mathbf{X} = \mathbf{X}^\star$.

Proof. First note that if $\mathbf{X}^\star$ is a critical point of f, then

$$\nabla f(\mathbf{X}^\star) = \mathbf{0}.$$

Now for any $\mathbf{X} \in \mathbb{R}^{n\times m}$ with $\mathrm{rank}(\mathbf{X}) \leq r$, the second order Taylor expansion gives

$$\begin{aligned} f(\mathbf{X}) =& f(\mathbf{X}^\star) + \langle \nabla f(\mathbf{X}^\star), \mathbf{X} - \mathbf{X}^\star\rangle + \frac{1}{2}[\nabla^2 f(\widetilde{\mathbf{X}})](\mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star) \\ =& f(\mathbf{X}^\star) + \frac{1}{2}[\nabla^2 f(\widetilde{\mathbf{X}})](\mathbf{X} - \mathbf{X}^\star, \mathbf{X} - \mathbf{X}^\star) \end{aligned}$$

where $\widetilde{\mathbf{X}} = t\mathbf{X}^\star + (1-t)\mathbf{X}$ for some $t \in [0, 1]$. This Taylor expansion together with (6.6) (both $\widetilde{\mathbf{X}}$ and $\mathbf{X}' - \mathbf{X}^\star$ have rank at most $2r$) gives

$$f(\mathbf{X}) - f(\mathbf{X}^\star) \geq a\|\mathbf{X} - \mathbf{X}^\star\|_F^2.$$

□

E.4 Proof of Lemma 6.3.1

Lemma E.4.1 (Lemma 6.3.1)**.** *Suppose $G(\mathbf{W})$ is defined as in* (6.9) *with $\mu > 0$. Then any critical point $\mathbf{W}$ of $G(\mathbf{W})$ belongs to $\mathcal{E}$, i.e.,*

$$\nabla G(\mathbf{W}) = \mathbf{0} \quad \Rightarrow \quad \mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}. \tag{6.11}$$

Proof. Any critical point (see Definition 6.2.1) $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ satisfies $\nabla G(\mathbf{W}) = \mathbf{0}$, i.e.,

$$\nabla f(\mathbf{U}\mathbf{V}^\top)\mathbf{V} + \mu\mathbf{U}\left(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right) = \mathbf{0}, \tag{E.17}$$

$$(\nabla f(\mathbf{U}\mathbf{V}^\top))^\top\mathbf{U} - \mu\mathbf{V}\left(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right) = \mathbf{0}. \tag{E.18}$$

By (E.18), we obtain

$$(\nabla f(\mathbf{U}\mathbf{V}^\top))^\top \mathbf{U} = \mu \left(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right)\mathbf{V}^\top.$$

Multiplying (E.17) by $\mathbf{U}^\top$ and plugging it in the expression for $\mathbf{U}^\top\nabla f(\mathbf{U}\mathbf{V}^\top)$ from the above equation gives

$$\left(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right)\mathbf{V}^\top\mathbf{V} + \mathbf{U}^\top\mathbf{U}\left(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right) = \mathbf{0},$$

which further implies

$$\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}\mathbf{V}^\top\mathbf{V}.$$

In order to show (6.11), note that $\mathbf{U}^\top\mathbf{U}$ and $\mathbf{V}^\top\mathbf{V}$ are the principal square roots (i.e., PSD square roots) of $\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top\mathbf{U}$ and $\mathbf{V}^\top\mathbf{V}\mathbf{V}^\top\mathbf{V}$, respectively. Utilizing the result that a PSD matrix $\mathbf{A}$ has a unique PSD matrix $\mathbf{B}$ such that $\mathbf{B}^k = \mathbf{A}$ for any $k \geq 1$ [271, Theorem 7.2.6], we obtain

$$\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}$$

for any critical point $\mathbf{W}$. □

E.5 Proof of Lemma E.1.1

Lemma E.5.1 (Lemma E.1.1)**.** *Suppose $g(\mathbf{W})$ is defined as in* (E.1) *with $\mu > 0$. Then we have*

$$g(\mathbf{W}) \geq \min\{\frac{\mu}{4}, \frac{1}{8}\} \left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2.$$

In particular, if we choose $\mu = \frac{1}{2}$, then we have

$$g(\mathbf{W}) = \frac{1}{8}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2 + \frac{1}{4}\left\|\mathbf{U}^\top\mathbf{U}^\star - \mathbf{V}^\top\mathbf{V}^\star\right\|_F^2.$$

Proof. We first rewrite the objective function $g(\mathbf{W})$:

$$\begin{aligned} g(\mathbf{W}) &= \frac{1}{2}\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{U}^\star\mathbf{V}^{\star\top}\right\|_F^2 + \frac{\mu}{4}\left\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right\|_F^2 \\ &\geq \min\{\mu, \frac{1}{2}\}\left(\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{U}^\star\mathbf{V}^{\star\top}\right\|_F^2 + \frac{1}{4}\left\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right\|_F^2\right) \\ &= \min\{\mu, \frac{1}{2}\}\left(\frac{1}{4}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2 + g'(\mathbf{W})\right), \end{aligned}$$

where the second line attains the equality when $\mu = \frac{1}{2}$, and $g'(\mathbf{W})$ in the last line is defined as

$$
\begin{aligned}
g'(\mathbf{W}) :=& \frac{1}{2}\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{U}^\star\mathbf{V}^{\star\top}\right\|_F^2 - \frac{1}{4}\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top}\right\|_F^2 \\
& - \frac{1}{4}\left\|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\right\|_F^2 + \frac{1}{4}\left\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right\|_F^2.
\end{aligned}
$$

We further show $g'(\mathbf{W})$ is always nonnegative:

$$
\begin{aligned}
g'(\mathbf{W}) &= \frac{1}{2}\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{U}^\star\mathbf{V}^{\star\top}\right\|_F^2 - \frac{1}{4}\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top}\right\|_F^2 \\
&\quad - \frac{1}{4}\left\|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\right\|_F^2 + \frac{1}{4}\left\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\right\|_F^2. \\
&= \frac{1}{2}\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{U}^\star\mathbf{V}^{\star\top}\right\|_F^2 + \frac{1}{2}\left\|\mathbf{U}^\top\mathbf{U}^\star\right\|_F^2 + \frac{1}{2}\left\|\mathbf{V}^\top\mathbf{V}^\star\right\|_F^2 \\
&\quad - \frac{1}{2}\operatorname{tr}\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^\top\mathbf{V}\right) - \frac{1}{4}\left\|\mathbf{U}^\star\mathbf{U}^{\star\top}\right\|_F^2 - \frac{1}{4}\left\|\mathbf{V}^\star\mathbf{V}^{\star\top}\right\|_F^2 \\
&= \frac{1}{2}\left\|\mathbf{U}^\top\mathbf{U}^\star - \mathbf{V}^\top\mathbf{V}^\star\right\|_F^2 + \frac{1}{2}\left\|\mathbf{U}^\star\mathbf{V}^{\star\top}\right\|_F^2 \\
&\quad - \frac{1}{4}\left\|\mathbf{U}^\star\mathbf{U}^{\star\top}\right\|_F^2 - \frac{1}{4}\left\|\mathbf{V}^\star\mathbf{V}^{\star\top}\right\|_F^2 \\
&= \frac{1}{2}\left\|\mathbf{U}^\top\mathbf{U}^\star - \mathbf{V}^\top\mathbf{V}^\star\right\|_F^2 \geq 0,
\end{aligned}
$$

where the last line follows because $\mathbf{U}^{\star\top}\mathbf{U}^\star = \mathbf{V}^{\star\top}\mathbf{V}^\star$. Thus, we have

$$
g(\mathbf{W}) \geq \min\{\frac{\mu}{4}, \frac{1}{8}\}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2,
$$

and

$$
g(\mathbf{W}) = \frac{1}{8}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2 + \frac{1}{4}\left\|\mathbf{U}^\top\mathbf{U}^\star - \mathbf{V}^\top\mathbf{V}^\star\right\|_F^2
$$

if $\mu = \frac{1}{2}$. □

E.6 Proof of Lemma E.1.2

Lemma E.6.1 (Lemma E.1.2)**.** *Let $\mathbf{X}^\star = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Psi}^\top = \sum_{i=1}^r \sigma_i\boldsymbol{\phi}_i\boldsymbol{\psi}_i^\top$ be a reduced SVD of $\mathbf{X}^\star$ and $g(\mathbf{W})$ be defined as in* (E.1) *with $\mu > 0$. Any $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ is a critical point of $g(\mathbf{W})$ if and only if $\mathbf{W} \in \mathcal{C}$ with*

$$
\mathcal{C} := \left\{\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi}\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{R} \in \mathcal{O}_r, \mathbf{\Lambda} \text{ is diagonal}, \mathbf{\Lambda} \geq \mathbf{0}, (\mathbf{\Sigma} - \mathbf{\Lambda})\mathbf{\Sigma} = \mathbf{0}\right\}. \tag{E.7}
$$

We first repeat that $\mathbf{X}^\star = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Psi}^\top =$ is a reduced SVD of $\mathbf{X}^\star$. We separate $\mathbf{U}$ into two parts—the projections onto the column space of $\mathbf{\Phi}$ and its orthogonal complement—by denoting $\mathbf{U} = \mathbf{\Phi}\mathbf{\Lambda}_1^{1/2}\mathbf{R}_1 + \mathbf{E}_1$ with $\mathbf{R}_1 \in \mathcal{O}_r$, $\mathbf{E}_1^\top\mathbf{\Phi} = \mathbf{0}$ and $\mathbf{\Lambda}_1$ being a $r \times r$ diagonal matrix with non-negative elements along its diagonal. Similarly, denote $\mathbf{V} = \mathbf{\Psi}\mathbf{\Lambda}_2^{1/2}\mathbf{R}_2 + \mathbf{E}_2$, where $\mathbf{R}_2 \in \mathcal{O}_r$, $\mathbf{E}_2^\top\mathbf{\Psi} = \mathbf{0}$, $\mathbf{\Lambda}_2$ is a $r \times r$ diagonal matrix with non-negative elements along

its diagonal. Recall that any critical point $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ satisfies

$$\nabla_{\mathbf{U}} \rho(\mathbf{U}, \mathbf{V}) = \mathbf{U}\mathbf{U}^\top \mathbf{U} - \mathbf{X}^\star \mathbf{V} = \mathbf{0},$$
$$\nabla_{\mathbf{V}} \rho(\mathbf{U}, \mathbf{V}) = \mathbf{V}\mathbf{V}^\top \mathbf{V} - \mathbf{X}^{\star\top} \mathbf{U} = \mathbf{0}.$$

Plugging $\mathbf{U} = \mathbf{\Phi}\mathbf{\Lambda}_1^{1/2}\mathbf{R}_1 + \mathbf{E}_1$ and $\mathbf{V} = \mathbf{\Psi}\mathbf{\Lambda}_2^{1/2}\mathbf{R}_2 + \mathbf{E}_2$ into the above equations gives

$$\mathbf{\Phi}\mathbf{\Lambda}_1^{3/2}\mathbf{R}_1 + \mathbf{\Phi}\mathbf{\Lambda}_1^{1/2}\mathbf{R}_1\mathbf{E}_1^\top\mathbf{E}_1 + \mathbf{E}_1\mathbf{R}_1^\top\mathbf{\Lambda}_1\mathbf{R}_1 + \mathbf{E}_1\mathbf{E}_1^\top\mathbf{E}_1 - \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Lambda}_2^{1/2}\mathbf{R}_2 = \mathbf{0}, \tag{E.19}$$

$$\mathbf{\Psi}\mathbf{\Lambda}_2^{3/2}\mathbf{R}_2 + \mathbf{\Psi}\mathbf{\Lambda}_2^{1/2}\mathbf{R}_2\mathbf{E}^\top\mathbf{E}_2 + \mathbf{E}_2\mathbf{R}_2^\top\mathbf{\Lambda}_2\mathbf{R}_2 + \mathbf{E}_2\mathbf{E}_2^\top\mathbf{E}_2 - \mathbf{\Psi}\mathbf{\Sigma}\mathbf{\Lambda}_1^{1/2}\mathbf{R}_1 = \mathbf{0}. \tag{E.20}$$

Since $\mathbf{E}_1$ is orthogonal to $\mathbf{\Phi}$, (E.19) further implies that

$$\mathbf{\Phi}\mathbf{\Lambda}_1^{3/2}\mathbf{R}_1 + \mathbf{\Phi}\mathbf{\Lambda}_1^{1/2}\mathbf{R}_1\mathbf{E}_1^\top\mathbf{E}_1 - \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Lambda}_2^{1/2}\mathbf{R}_2 = \mathbf{0}, \tag{E.21}$$
$$\mathbf{E}_1\mathbf{R}_1^\top\mathbf{\Lambda}_1\mathbf{R}_1 + \mathbf{E}_1\mathbf{E}_1^\top\mathbf{E}_1 = \mathbf{0}. \tag{E.22}$$

From (E.22), we have

$$\begin{aligned}&\left\langle \mathbf{E}_1\mathbf{R}_1^\top\mathbf{\Lambda}_1\mathbf{R}_1 + \mathbf{E}_1\mathbf{E}_1^\top\mathbf{E}_1, \mathbf{E}_1 \right\rangle \\ &= \left\langle \mathbf{R}_1^\top\mathbf{\Lambda}_1\mathbf{R}_1, \mathbf{E}_1^\top\mathbf{E}_1 \right\rangle + \|\mathbf{E}_1\|_F^2 = 0,\end{aligned}$$

which further implies $\|\mathbf{E}_1\|_F^2 = 0$ by noting that $\left\langle \mathbf{R}_1^\top\mathbf{\Lambda}_1\mathbf{R}_1, \mathbf{E}_1^\top\mathbf{E}_1 \right\rangle \geq 0$ since it is the inner product between two PSD matrices. Thus $\mathbf{E}_1 = \mathbf{0}$. With a similar argument we also have $\mathbf{E}_2 = \mathbf{0}$.

With $\mathbf{E}_1 = \mathbf{E}_2 = \mathbf{0}$, (E.21) reduces to

$$\mathbf{\Phi}\mathbf{\Lambda}_1^{3/2}\mathbf{R}_1 - \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Lambda}_2^{1/2}\mathbf{R}_2 = \mathbf{0}.$$

Since $\mathbf{\Phi}$ is orthogonal and $\mathbf{R}_1 \in \mathcal{O}_r$, the above equation implies that

$$\mathbf{\Lambda}_1^{3/2} = \mathbf{\Sigma}\mathbf{\Lambda}_2^{1/2}\mathbf{R}_2\mathbf{R}_1^\top.$$

Let Ω denote the set of locations of the non-zero diagonals in $\mathbf{\Lambda}_2$, i.e., $\mathbf{\Lambda}_2[i,i] > 0$ for all $i \in \Omega$. Then $[\mathbf{R}_1^\top]_\Omega = [\mathbf{R}_2^\top]_\Omega$ since otherwise $\mathbf{\Sigma}\mathbf{\Lambda}_2^{1/2}\mathbf{R}_2\mathbf{R}_1^\top$ is not a diagonal matrix anymore. Then we have

$$\mathbf{\Lambda}_1^{3/2} = \mathbf{\Sigma}\mathbf{\Lambda}_2^{1/2} \tag{E.23}$$

implying that the set of the locations of non-zero diagonals in $\mathbf{\Lambda}_1$ is identical to Ω. A similar argument applied to (E.20) gives

$$\mathbf{\Lambda}_2^{3/2} = \mathbf{\Sigma}\mathbf{\Lambda}_1^{1/2}. \tag{E.24}$$

Noting that (E.23) implies $\mathbf{\Lambda}_1^{3/2}[i,i] = \mathbf{\Sigma}[i,i]\mathbf{\Lambda}_2^{1/2}[i,i]$ and (E.24) implies $\mathbf{\Lambda}_2^{3/2}[i,i] = \mathbf{\Sigma}[i,i]\mathbf{\Lambda}_1^{1/2}[i,i]$, for all $i \in \Omega$ we have $\mathbf{\Lambda}_1[i,i] = \mathbf{\Lambda}_2[i,i] = \mathbf{\Sigma}[i,i]$. For $i \notin \Omega$, we have $\mathbf{\Lambda}_1[i,i] = \mathbf{\Lambda}_2[i,i] = 0$. Thus $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2$. For convenience, denote $\mathbf{\Lambda} = \mathbf{\Lambda}_1 = \mathbf{\Lambda}_2$ with $\mathbf{\Lambda}[i,i] = \lambda_i$.

Finally, we note that $\mathbf{U} = \mathbf{\Phi}\mathbf{\Lambda}^{1/2}\mathbf{R}_1 = \sum_{i\in\Omega} \lambda_i \boldsymbol{\phi}_i \mathbf{R}_1[i,:]$ and $\mathbf{V} = \mathbf{\Psi}\mathbf{\Lambda}^{1/2}\mathbf{R}_2 = \sum_{i\in\Omega} \lambda_i \boldsymbol{\psi}_i \mathbf{R}_2[i,:]$ implying that only $[\mathbf{R}_1^\top]_\Omega$ and $[\mathbf{R}_2^\top]_\Omega$ play a role in $\mathbf{U}$ and $\mathbf{V}$, respectively. Thus one can set $\mathbf{R}_1 = \mathbf{R}_2$ since we already proved $[\mathbf{R}_1^\top]_\Omega = [\mathbf{R}_2^\top]_\Omega$.

E.7 Proof of Lemma E.1.3

Lemma E.7.1 (Lemma E.1.3)**.** *For any* $\mathbf{\Delta} = \begin{bmatrix} \mathbf{\Delta}_\mathbf{U} \\ \mathbf{\Delta}_\mathbf{V} \end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$ *and* $\mathbf{W} \in \mathcal{E}$ *where* $\mathcal{E}$ *is defined in* (6.10)*, we have*

$$\|\mathbf{\Delta}_\mathbf{U}\mathbf{U}^\top\|_F^2 + \|\mathbf{\Delta}_\mathbf{V}\mathbf{V}^\top\|_F^2 = \|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\|_F^2 + \|\mathbf{\Delta}_\mathbf{V}\mathbf{U}^\top\|_F^2, \tag{E.9}$$

and

$$\nabla^2\rho(\mathbf{W}) \succeq \mathbf{0}. \tag{E.10}$$

Proof. Utilizing the result that any point $\mathbf{W} \in \mathcal{E}$ satisfies $\widehat{\mathbf{W}}^\top\mathbf{W} = \mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V} = \mathbf{0}$, we directly obtain

$$\|\mathbf{\Delta}_\mathbf{U}\mathbf{U}^\top\|_F^2 + \|\mathbf{\Delta}_\mathbf{V}\mathbf{V}^\top\|_F^2 = \|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\|_F^2 + \|\mathbf{\Delta}_\mathbf{V}\mathbf{U}^\top\|_F^2$$

since $\|\mathbf{\Delta}_\mathbf{U}\mathbf{U}^\top\|_F^2 = \operatorname{tr}\left(\mathbf{\Delta}_\mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{\Delta}_\mathbf{U}\right) = \operatorname{tr}\left(\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\mathbf{V}\mathbf{\Delta}_\mathbf{U}\right) = \|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\|_F^2$ (and similarly for the other two terms).

We then rewrite the last two terms in (E.4) as

$$\begin{aligned}
\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{\Delta}}^\top, \mathbf{\Delta}\mathbf{W}^\top \right\rangle + \left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \mathbf{\Delta}\mathbf{\Delta}^\top \right\rangle &= \left\langle \widehat{\mathbf{W}}^\top\mathbf{\Delta}, \mathbf{\Delta}^\top\widehat{\mathbf{W}} \right\rangle + \left\langle \widehat{\mathbf{W}}^\top\mathbf{\Delta}, \widehat{\mathbf{W}}^\top\mathbf{\Delta} \right\rangle \\
&= \left\langle \widehat{\mathbf{W}}^\top\mathbf{\Delta}, \widehat{\mathbf{W}}^\top\mathbf{\Delta} + \mathbf{\Delta}^\top\widehat{\mathbf{W}} \right\rangle \\
&= \frac{1}{2}\left\langle \widehat{\mathbf{W}}^\top\mathbf{\Delta} + \mathbf{\Delta}^\top\widehat{\mathbf{W}}, \widehat{\mathbf{W}}^\top\mathbf{\Delta} + \mathbf{\Delta}^\top\widehat{\mathbf{W}} \right\rangle \\
&\quad + \frac{1}{2}\left\langle \widehat{\mathbf{W}}^\top\mathbf{\Delta} - \mathbf{\Delta}^\top\widehat{\mathbf{W}}, \widehat{\mathbf{W}}^\top\mathbf{\Delta} + \mathbf{\Delta}^\top\widehat{\mathbf{W}} \right\rangle \\
&= \frac{1}{2}\left\| \widehat{\mathbf{W}}^\top\mathbf{\Delta} + \mathbf{\Delta}^\top\widehat{\mathbf{W}} \right\|_F^2
\end{aligned}$$

where the last line holds because $\left\langle \mathbf{A} - \mathbf{A}^\top, \mathbf{A} + \mathbf{A}^\top \right\rangle = 0$. Plugging these with the factor $\widehat{\mathbf{W}}^\top\mathbf{W} = \mathbf{0}$ into the Hessian quadrature form $[\nabla^2\rho(\mathbf{W})](\mathbf{\Delta}, \mathbf{\Delta})$ defined in (E.4) gives

$$[\nabla^2\rho(\mathbf{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}) \geq \frac{\mu}{2}\left\|\widehat{\mathbf{W}}^\top\boldsymbol{\Delta}+\boldsymbol{\Delta}^\top\widehat{\mathbf{W}}\right\|_F^2 \geq 0.$$

This implies that the Hessian of ρ evaluated at any $\mathbf{W}\in\mathcal{E}$ is PSD, i.e., $\nabla^2\rho(\mathbf{W})\succeq\mathbf{0}$.[46] □

E.8 Proof of Theorem E.1.1 (strict saddle property for (E.1))

Theorem E.8.1 (Theorem E.1.1)**.** *Let $g(\mathbf{W})$ be defined as in* (E.1) *with $\mu>0$ and $\mathrm{rank}(\mathbf{X}^\star)=r$. Let $\mathbf{W}=\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ be any critical point satisfying $\nabla g(\mathbf{W})=\mathbf{0}$, i.e., $\mathbf{W}\in\mathcal{C}$. Any $\mathbf{W}\in\mathcal{C}\setminus\mathcal{X}$ is a strict saddle of $g(\mathbf{W})$ satisfying*

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -\frac{1}{2}\left\|\mathbf{W}\mathbf{W}^\top-\mathbf{W}^\star\mathbf{W}^{\star\top}\right\| \leq -\sigma_r(\mathbf{X}^\star). \tag{E.11}$$

Furthermore, $g(\mathbf{W})$ is not strongly convex at any global minimum point $\mathbf{W}\in\mathcal{X}$.

Proof. We begin the proof of Theorem E.1.1 by characterizing any $\mathbf{W}\in\mathcal{C}\setminus\mathcal{X}$. For this purpose, let $\mathbf{W}=\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$, where $\mathbf{U}=\boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2}\mathbf{R},\mathbf{V}=\boldsymbol{\Psi}\boldsymbol{\Lambda}^{1/2}\mathbf{R},\mathbf{R}\in\mathcal{O}_r,\boldsymbol{\Lambda}$ is diagonal, $\boldsymbol{\Lambda}\geq\mathbf{0},(\boldsymbol{\Sigma}-\boldsymbol{\Lambda})\boldsymbol{\Sigma}=\mathbf{0}$, and $\mathrm{rank}(\boldsymbol{\Lambda})<r$. Denote the corresponding optimal solution $\mathbf{W}^\star=\begin{bmatrix}\mathbf{U}^\star\\ \mathbf{V}^\star\end{bmatrix}$, where $\mathbf{U}^\star=\boldsymbol{\Phi}\boldsymbol{\Sigma}^{1/2}\mathbf{R},\mathbf{V}^\star=\boldsymbol{\Psi}\boldsymbol{\Sigma}^{1/2}\mathbf{R}$. Let

$$k=\arg\max_i \sigma_i-\lambda_i$$

denote the location of the first zero diagonal element in $\boldsymbol{\Lambda}$. Noting that $\lambda_i\in\{\sigma_i,0\}$, we conclude that

$$\lambda_k=0,\quad \boldsymbol{\phi}_k^\top\mathbf{U}=\mathbf{0},\quad \boldsymbol{\psi}_k^\top\mathbf{V}=\mathbf{0}. \tag{E.25}$$

In words, $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$ are orthogonal to $\mathbf{U}$ and $\mathbf{V}$, respectively. Let $\boldsymbol{\alpha}\in\mathbb{R}^r$ be the eigenvector associated with the smallest eigenvalue of $\mathbf{W}^\top\mathbf{W}$. Such $\boldsymbol{\alpha}$ simultaneously lives in the null spaces of $\mathbf{U}$ and $\mathbf{V}$ since $\mathbf{W}$ is rank deficient indicating

$$0=\boldsymbol{\alpha}^\top\mathbf{W}^\top\mathbf{W}\boldsymbol{\alpha}=\boldsymbol{\alpha}^\top\mathbf{U}^\top\mathbf{U}\boldsymbol{\alpha}+\boldsymbol{\alpha}^\top\mathbf{V}^\top\mathbf{V}\boldsymbol{\alpha},$$

which further implies

$$\begin{cases}\boldsymbol{\alpha}^\top\mathbf{U}^\top\mathbf{U}\boldsymbol{\alpha}=0,\\ \boldsymbol{\alpha}^\top\mathbf{V}^\top\mathbf{V}\boldsymbol{\alpha}=0.\end{cases} \tag{E.26}$$

With this property, we construct $\boldsymbol{\Delta}$ by setting $\boldsymbol{\Delta}_{\mathbf{U}}=\boldsymbol{\phi}_k\boldsymbol{\alpha}^\top$ and $\boldsymbol{\Delta}_{\mathbf{V}}=\boldsymbol{\psi}_k\boldsymbol{\alpha}^\top$. Now we show that $\mathbf{W}$ is a strict saddle by arguing that $g(\mathbf{W})$ has a strictly negative curvature along the constructed direction $\boldsymbol{\Delta}$, i.e., $[\nabla^2 g(\mathbf{W})](\boldsymbol{\Delta},\boldsymbol{\Delta})<0$.

[46] This can also be observed since any critical point $\mathbf{W}$ is a global minimum of $\rho(\mathbf{W})$, which directly indicates that $\nabla^2\rho(\mathbf{W})\succeq\mathbf{0}$.

To that end, we compute the five terms in (E.3) as follows

$$\begin{aligned}
\left\|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top + \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\|_F^2 &= 0 \quad (\text{since} \quad (E.26)),\\
\left\langle \mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star, \boldsymbol{\Delta}_{\mathbf{U}}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\rangle &= \lambda_k - \sigma_k = -\sigma_k \quad (\text{since} \quad (E.25)),\\
\left\langle \widehat{\mathbf{W}}^\top\mathbf{W}, \widehat{\boldsymbol{\Delta}}^\top\boldsymbol{\Delta}\right\rangle &= 0 \quad (\text{since} \quad \widehat{\mathbf{W}}^\top\mathbf{W} = \mathbf{0}),\\
\left\langle \widehat{\mathbf{W}}\widehat{\boldsymbol{\Delta}}^\top, \boldsymbol{\Delta}\mathbf{W}^\top\right\rangle &= \operatorname{tr}\left(\widehat{\boldsymbol{\Delta}}^\top\mathbf{W}\boldsymbol{\Delta}^T\widehat{\mathbf{W}}\right) = 0,\\
\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right\rangle &= \operatorname{tr}\left(\widehat{\mathbf{W}}^\top\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\widehat{\mathbf{W}}\right) = 0,
\end{aligned}$$

where $\widehat{\mathbf{W}}^\top\mathbf{W} = \mathbf{0}$ since $\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V} = \mathbf{0}$, the last two lines utilize $\widehat{\boldsymbol{\Delta}}^\top\mathbf{W} = \mathbf{0}$ (or $\widehat{\mathbf{W}}^\top\boldsymbol{\Delta} = \mathbf{0}$) because $\widehat{\boldsymbol{\Delta}}^\top\mathbf{W} = \boldsymbol{\alpha}\boldsymbol{\phi}_k^\top\mathbf{U} - \boldsymbol{\alpha}\boldsymbol{\psi}_k^\top\mathbf{V} = \mathbf{0}$ (see (E.25)). Plugging these terms into (E.3) gives

$$\begin{aligned}
[\nabla^2 g(\mathbf{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}) &= \left\|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top + \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\|_F^2 + 2\left\langle \mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star, \boldsymbol{\Delta}_{\mathbf{U}}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\rangle\\
&\quad + \mu\left\langle \widehat{\mathbf{W}}^\top\mathbf{W}, \widehat{\boldsymbol{\Delta}}^\top\boldsymbol{\Delta}\right\rangle + \mu\left\langle \widehat{\mathbf{W}}\widehat{\boldsymbol{\Delta}}^\top, \boldsymbol{\Delta}\mathbf{W}^\top\right\rangle\\
&\quad + \mu\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right\rangle\\
&= -2\sigma_k.
\end{aligned}$$

The proof of the strict saddle property is completed by noting that

$$\|\boldsymbol{\Delta}\|_F^2 = \|\boldsymbol{\Delta}_{\mathbf{U}}\|_F^2 + \|\boldsymbol{\Delta}_{\mathbf{V}}\|_F^2 = \left\|\boldsymbol{\phi}_k\boldsymbol{\alpha}^\top\right\|_F^2 + \left\|\boldsymbol{\psi}_k\boldsymbol{\alpha}^\top\right\|_F^2 = 2,$$

which further implies

$$\begin{aligned}
\lambda_{\min}\left(\nabla^2 g(\mathbf{W})\right) &\le \frac{[\nabla^2 g(\mathbf{W})](\boldsymbol{\Delta},\boldsymbol{\Delta})}{\|\boldsymbol{\Delta}\|_F^2} \le -\frac{2\sigma_k}{2}\\
&= -\|\boldsymbol{\Lambda} - \boldsymbol{\Sigma}\| = -\frac{1}{2}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|,
\end{aligned}$$

where the first equality holds because

$$\|\boldsymbol{\Lambda} - \boldsymbol{\Sigma}\| = \max_i \sigma_i - \lambda_i = \sigma_k,$$

and the second equality follows since

$$\begin{aligned}
&\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top} = \frac{1}{2}\mathbf{Q}\left(\boldsymbol{\Lambda} - \boldsymbol{\Sigma}\right)\mathbf{Q}^\top,\\
&\mathbf{Q} = \begin{bmatrix}\boldsymbol{\Phi}/\sqrt{2}\\ \boldsymbol{\Psi}/\sqrt{2}\end{bmatrix}, \quad \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}.
\end{aligned}$$

We finish the proof of (E.11) by noting that

$$\sigma_k = \sigma_k(\mathbf{X}^\star) \ge \sigma_r(\mathbf{X}^\star).$$

Now suppose $\mathbf{W}^\star \in \mathcal{X}$. Applying (E.10), which states that the Hessian of ρ evaluated at any critical point $\mathbf{W}$ is PSD, we have

$$\begin{aligned}[\nabla^2 g(\mathbf{W}^\star)][\mathbf{\Delta}, \mathbf{\Delta}] &= \left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^{\star\top} + \mathbf{U}^\star\mathbf{\Delta}_\mathbf{V}^\top\right\|_F^2 + 2\left\langle \mathbf{U}^\star\mathbf{V}^{\star\top} - \mathbf{X}^\star, \mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\rangle \\ &\quad + [\nabla^2\rho(\mathbf{W}^\star)][\mathbf{\Delta}, \mathbf{\Delta}] \\ &\geq \left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^{\star\top} + \mathbf{U}^\star\mathbf{\Delta}_\mathbf{V}^\top\right\|_F^2 + 2\left\langle \mathbf{U}^\star\mathbf{V}^{\star\top} - \mathbf{X}^\star, \mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\rangle \\ &\geq 0\end{aligned}$$

since $\mathbf{U}^\star\mathbf{V}^{\star\top} - \mathbf{X}^\star = \mathbf{0}$. We show g is not strongly convex at $\mathbf{W}^\star$ by arguing that $\lambda_{\min}(\nabla^2 g(\mathbf{W}^\star)) = 0$. For this purpose, we first recall that $\mathbf{U}^\star = \mathbf{\Phi}\mathbf{\Sigma}^{1/2}, \mathbf{V}^\star = \mathbf{\Psi}\mathbf{\Sigma}^{1/2}$, where we assume $\mathbf{R} = \mathbf{I}$ without loss of generality. Let $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_r\}$ be the standard orthobasis for $\mathbb{R}^r$, i.e., $\mathbf{e}_\ell$ is the ℓ-th column of the $r \times r$ identity matrix. Construct $\mathbf{\Delta}_{(i,j)} = \begin{bmatrix}\mathbf{\Delta}_\mathbf{U}^{(i,j)} \\ \mathbf{\Delta}_\mathbf{V}^{(i,j)}\end{bmatrix}$, where

$$\mathbf{\Delta}_\mathbf{U}^{(i,j)} = \mathbf{U}^\star\mathbf{e}_j\mathbf{e}_i^\top - \mathbf{U}^\star\mathbf{e}_i\mathbf{e}_j^\top, \quad \mathbf{\Delta}_\mathbf{V}^{(i,j)} = \mathbf{V}^\star\mathbf{e}_j\mathbf{e}_i^\top - \mathbf{U}^\star\mathbf{e}_i\mathbf{e}_j^\top,$$

for any $1 \leq i < j \leq r$. That is, the ℓ-th columns of the matrices $\mathbf{\Delta}_\mathbf{U}^{(i,j)}$ and $\mathbf{\Delta}_\mathbf{V}^{(i,j)}$ are respectively given by

$$\mathbf{\Delta}_\mathbf{U}^{(i,j)}[:,\ell] = \begin{cases}\sigma_j^{1/2}\boldsymbol{\phi}_j, & \ell = i, \\ -\sigma_i^{1/2}\boldsymbol{\phi}_i, & \ell = j, \\ \mathbf{0}, & \text{otherwise,}\end{cases},$$

$$\mathbf{\Delta}_\mathbf{V}^{(i,j)}[:,\ell] = \begin{cases}\sigma_j^{1/2}\boldsymbol{\psi}_j, & \ell = i, \\ -\sigma_i^{1/2}\boldsymbol{\psi}_i, & \ell = j, \\ \mathbf{0}, & \text{otherwise,}\end{cases}$$

for any $1 \leq i < j \leq r$. We then compute the five terms in (E.3) as follows

$$\left\|\mathbf{\Delta}_\mathbf{U}^{(i,j)}\mathbf{V}^{\star\top} + \mathbf{U}^\star(\mathbf{\Delta}_\mathbf{V}^{(i,j)})^\top\right\|_F^2 = \left\|\sigma_i^{1/2}\sigma_j^{1/2}\left(\boldsymbol{\phi}_j\boldsymbol{\psi}_i^\top - \boldsymbol{\phi}_i\boldsymbol{\psi}_j^\top + \boldsymbol{\phi}_i\boldsymbol{\psi}_j^\top - \boldsymbol{\phi}_j\boldsymbol{\psi}_i^\top\right)\right\|_F^2 = 0,$$

$$\begin{aligned}&\langle \mathbf{U}^\star\mathbf{V}^{\star\top} - \mathbf{X}^\star, \mathbf{\Delta}_\mathbf{U}^{(i,j)}(\mathbf{\Delta}_\mathbf{V}^{(i,j)})^\top\rangle = 0 \text{ (as } \mathbf{U}^\star\mathbf{V}^{\star\top} - \mathbf{X}^\star = \mathbf{0}), \\ &\langle \widehat{\mathbf{W}}^{\star\top}\mathbf{W}^\star, \widehat{\mathbf{\Delta}}_{(i,j)}^\top\mathbf{\Delta}_{(i,j)}\rangle = 0 \quad (\text{as } \widehat{\mathbf{W}}^{\star\top}\mathbf{W}^\star = \mathbf{0}), \\ &\langle \widehat{\mathbf{W}}^\star\widehat{\mathbf{\Delta}}_{(i,j)}^\top, \mathbf{\Delta}_{(i,j)}\mathbf{W}^{\star\top}\rangle = \mathrm{tr}(\widehat{\mathbf{W}}^{\star\top}\mathbf{\Delta}_{(i,j)}\mathbf{W}^{\star\top}\widehat{\mathbf{\Delta}}_{(i,j)}) = 0, \\ &\langle \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \mathbf{\Delta}_{(i,j)}\mathbf{\Delta}_{(i,j)}^\top\rangle = \mathrm{tr}(\widehat{\mathbf{W}}^{\star\top}\mathbf{\Delta}_{(i,j)}\mathbf{\Delta}_{(i,j)}^\top\widehat{\mathbf{W}}^\star) = 0,\end{aligned}$$

where the last two lines hold because

$$\widehat{\mathbf{W}}^{\star\top}\mathbf{\Delta}_{(i,j)} = \mathbf{U}^{\star\top}\mathbf{U}^\star(\mathbf{e}_j\mathbf{e}_i^\top - \mathbf{e}_i\mathbf{e}_j^\top) - \mathbf{V}^{\star\top}\mathbf{V}^\star(\mathbf{e}_j\mathbf{e}_i^\top - \mathbf{e}_i\mathbf{e}_j^\top) = \mathbf{0}$$

since $\mathbf{U}^{\star\top}\mathbf{U}^\star = \mathbf{V}^{\star\top}\mathbf{V}^\star$.

Thus, we obtain the Hessian evaluated at the optimal solution point $\mathbf{W}^\star$ along the direction $\mathbf{\Delta}^{(i,j)}$:

$$\left[\nabla^2 g(\mathbf{W}^\star)\right]\left(\mathbf{\Delta}^{(i,j)}, \mathbf{\Delta}^{(i,j)}\right) = 0$$

for all $1 \leq i < j \leq r$. This proves that $g(\mathbf{W})$ is not strongly convex at a global minimum point $\mathbf{W}^\star \in \mathcal{X}$. □

E.9 Proof of Theorem E.1.2 (strict saddle property of $g(\mathbf{W})$ when over-parameterized)

Theorem E.9.1 (Theorem E.1.2)**.** *Let $\mathbf{X}^\star = \mathbf{\Phi\Sigma\Psi}^\top = \sum_{i=1}^{r'} \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^\top$ be a reduced SVD of $\mathbf{X}^\star$ with $r' \leq r$, and let $g(\mathbf{W})$ be defined as in* (E.1) *with $\mu > 0$. Any $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ is a critical point of $g(\mathbf{W})$ if and only if $\mathbf{W} \in \mathcal{C}$ with*

$$\mathcal{C} := \left\{\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi\Lambda}^{1/2}\mathbf{R}, \mathbf{R}\mathbf{R}^\top = \mathbf{I}_{r'}, \mathbf{\Lambda} \text{ is diagonal}, \mathbf{\Lambda} \geq \mathbf{0}, (\mathbf{\Sigma} - \mathbf{\Lambda})\mathbf{\Sigma} = \mathbf{0}\right\}$$

Further, all the local minima (which are also global) belong to the following set

$$\mathcal{X} = \left\{\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi\Sigma}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi\Sigma}^{1/2}\mathbf{R}, \mathbf{R}\mathbf{R}^\top = \mathbf{I}_{r'}\right\}$$

Finally, any $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$ is a strict saddle of $g(\mathbf{W})$ satisfying

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -\frac{1}{2}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\| \leq -\sigma_{r'}(\mathbf{X}^\star).$$

Proof. Let $\mathbf{X}^\star = \mathbf{\Phi\Sigma\Psi}^\top = \sum_{i=1}^{r'} \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^\top$ be a reduced SVD of $\mathbf{X}^\star$ with $r' \leq r$. Using an approach similar to that in Appendix E.6 for proving Lemma E.1.2, we can show that any $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ is a critical point of $g(\mathbf{W})$ if and only if $\mathbf{W} \in \mathcal{C}$ with

$$\mathcal{C} = \left\{\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi\Lambda}^{1/2}\mathbf{R}, \mathbf{R}\mathbf{R}^\top = \mathbf{I}_{r'}, \mathbf{\Lambda} \text{ is diagonal}, \mathbf{\Lambda} \geq \mathbf{0}, (\mathbf{\Sigma} - \mathbf{\Lambda})\mathbf{\Sigma} = \mathbf{0}\right\}.$$

Recall that

$$\mathcal{X} = \left\{\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi\Sigma}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi\Sigma}^{1/2}\mathbf{R}, \mathbf{R}\mathbf{R}^\top = \mathbf{I}_{r'}\right\}.$$

It is clear that $\mathcal{X}$ is the set of optimal solutions since for any $\mathbf{W} \in \mathcal{X}$, $g(\mathbf{W})$ achieves its global minimum, i.e., $g(\mathbf{W}) = 0$.

Using an approach similar to that in Appendix E.8 for proving Theorem E.1.1, we can show that any $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$ is a strict saddle satisfying

$$\lambda_{\min}\left(\nabla^2 g(\mathbf{W})\right) \leq -\sigma_{r'}(\mathbf{X}^\star).$$

□

E.10 Proof of Theorem E.1.3 (strict saddle property of $g(\mathbf{W})$ when under-parameterized)

Theorem E.10.1 (Theorem E.1.3)**.** *Let $\mathbf{X}^\star = \mathbf{\Phi\Sigma\Psi}^\top = \sum_{i=1}^{r'} \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^\top$ be a reduced SVD of $\mathbf{X}^\star$ with $r' > r$ and $\sigma_r(\mathbf{X}^\star) > \sigma_{r+1}(\mathbf{X}^\star)$. Also let $g(\mathbf{W})$ be defined as in* (E.1) *with $\mu > 0$. Any $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ is a critical point of $g(\mathbf{W})$ if and only if $\mathbf{W} \in \mathcal{C}$ with*

$$\mathcal{C} := \Big\{ \mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi}[:,\Omega]\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}[:,\Omega]\mathbf{\Lambda}^{1/2}\mathbf{R}, \\ \mathbf{\Lambda} = \mathbf{\Sigma}[\Omega,\Omega], \mathbf{R}\mathbf{R}^\top = \mathbf{I}_\ell, \Omega \subset \{1,2,\ldots,r'\}, |\Omega| = \ell \le r \Big\}$$

where we recall that $\mathbf{\Phi}[:,\Omega]$ is a submatrix of $\mathbf{\Phi}$ obtained by keeping the columns indexed by Ω and $\mathbf{\Sigma}[\Omega,\Omega]$ is an $\ell \times \ell$ matrix obtained by taking the elements of $\mathbf{\Sigma}$ in rows and columns indexed by Ω.

Further, all local minima belong to the following set

$$\mathcal{X} = \Big\{ \mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{\Lambda} = \mathbf{\Sigma}[1:r,1:r], \mathbf{R} \in \mathcal{O}_r, \mathbf{U} = \mathbf{\Phi}[:,1:r]\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}[:,1:r]\mathbf{\Lambda}^{1/2}\mathbf{R} \Big\}.$$

Finally, any $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$ is a strict saddle of $g(\mathbf{W})$ satisfying

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \le -(\sigma_r(\mathbf{X}^\star) - \sigma_{r+1}(\mathbf{X}^\star)).$$

Proof. Let $\mathbf{X}^\star = \mathbf{\Phi\Sigma\Psi}^\top = \sum_{i=1}^{r'} \sigma_i \boldsymbol{\phi}_i \boldsymbol{\psi}_i^\top$ be a reduced SVD of $\mathbf{X}^\star$ with $r' > r$ and $\sigma_r(\mathbf{X}^\star) > \sigma_{r+1}(\mathbf{X}^\star)$. Using an approach similar to that in Appendix E.6 for proving Lemma E.1.2, we can show that any $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ is a critical point of $g(\mathbf{W})$ if and only if $\mathbf{W} \in \mathcal{C}$ with

$$\mathcal{C} = \Big\{ \mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi}[:,\Omega]\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}[:,\Omega]\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{\Lambda} = \mathbf{\Sigma}[\Omega,\Omega], \mathbf{R}\mathbf{R}^\top = \mathbf{I}_\ell, \Omega \subset \{1,\ldots,r'\}, |\Omega| = \ell \le r \Big\}.$$

Intuitively, a critical point is one such that $\mathbf{U}\mathbf{V}^\top$ is a rank-ℓ approximation to $\mathbf{X}^\star$ with $\ell \le r$ and $\mathbf{U}$ and $\mathbf{V}$ are equal factors of their product $\mathbf{U}\mathbf{V}^\top$.

It follows from the Eckart-Young-Mirsky theorem [271] that the set of optimal solutions is given by

$$\mathcal{X} = \Big\{ \mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} : \mathbf{U} = \mathbf{\Phi}[:,1:r]\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}[:,1:r]\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{\Lambda} = \mathbf{\Sigma}[1:r,1:r], \mathbf{R} \in \mathcal{O}_r \Big\}.$$

Now we characterize any $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$ by letting $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$, where

$$
\begin{aligned}
&\mathbf{U} = \mathbf{\Phi}[:, \Omega]\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}[:, \Omega]\mathbf{\Lambda}^{1/2}\mathbf{R}, \\
&\mathbf{\Lambda} = \mathbf{\Sigma}[\Omega, \Omega], \mathbf{R} \in \mathbb{R}^{\ell \times r}, \mathbf{R}\mathbf{R}^\top = \mathbf{I}_\ell, \\
&\Omega \subset \{1, 2, \ldots, r'\}, |\Omega| = \ell \leq r, \Omega \neq \{1, 2, \ldots, r\}.
\end{aligned}
$$

Let $\boldsymbol{\alpha} \in \mathbb{R}^r$ be the eigenvector associated with the smallest eigenvalue of $\mathbf{U}^\top\mathbf{U}$ (or $\mathbf{V}^\top\mathbf{V}$). By the typical structures in $\mathbf{U}$ and $\mathbf{V}$ (see the above equation), we have

$$
\begin{aligned}
\|\mathbf{V}\boldsymbol{\alpha}\|_F^2 &= \|\mathbf{U}\boldsymbol{\alpha}\|_F^2 = \sigma_r^2(\mathbf{U}) \\
&= \begin{cases} \sigma_j(\mathbf{X}^\star), & |\Omega| = r \text{ and } j = \max\Omega \\ 0, & |\Omega| < r, \end{cases}
\end{aligned} \tag{E.27}
$$

where $j > r$ because $\Omega \neq \{1, 2, \ldots, r\}$. Note that there always exists an index

$$
i \in \{1, 2, \ldots, r\}, i \notin \Omega
$$

since $\Omega \neq \{1, 2, \ldots, r\}$ and $|\Omega| \leq r$. We construct $\mathbf{\Delta}$ by setting

$$
\mathbf{\Delta}_\mathbf{U} = \boldsymbol{\phi}_i\boldsymbol{\alpha}^\top, \quad \mathbf{\Delta}_\mathbf{V} = \boldsymbol{\psi}_i\boldsymbol{\alpha}^\top.
$$

Since $i \notin \Omega$, we have

$$
\begin{aligned}
\mathbf{U}^\top\mathbf{\Delta}_\mathbf{U} = \mathbf{U}^\top\boldsymbol{\phi}_i\boldsymbol{\alpha}^\top = \mathbf{0}, \\
\mathbf{V}^\top\mathbf{\Delta}_\mathbf{V} = \mathbf{V}^\top\boldsymbol{\psi}_i\boldsymbol{\alpha}^\top = \mathbf{0}.
\end{aligned} \tag{E.28}
$$

We compute the five terms in (E.3) as follows

$$
\left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F^2 = \left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\right\|_F^2 + \left\|\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F^2 + 2\operatorname{tr}\left(\mathbf{U}^\top\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\mathbf{\Delta}_\mathbf{V}\right) = 2\sigma_r^2(\mathbf{U}),
$$

$$
\begin{aligned}
\left\langle \mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star, \mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top \right\rangle &= \left\langle \mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star, \boldsymbol{\phi}_i\boldsymbol{\psi}_i^\top \right\rangle = -\left\langle \mathbf{X}^\star, \boldsymbol{\phi}_i\boldsymbol{\psi}_i^\top \right\rangle = -\sigma_i(\mathbf{X}^\star), \\
\left\langle \widehat{\mathbf{W}}^\top\mathbf{W}, \widehat{\mathbf{\Delta}}^\top\mathbf{\Delta} \right\rangle &= 0 \quad (\text{since} \quad \widehat{\mathbf{W}}^\top\mathbf{W} = \mathbf{0}), \\
\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{\Delta}}^\top, \mathbf{\Delta}\mathbf{W}^\top \right\rangle &= \operatorname{tr}\left(\widehat{\mathbf{W}}^\top\mathbf{\Delta}\mathbf{W}^\top\widehat{\mathbf{\Delta}}\right) = 0, \\
\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \mathbf{\Delta}\mathbf{\Delta}^\top \right\rangle &= \operatorname{tr}\left(\widehat{\mathbf{W}}^\top\mathbf{\Delta}\mathbf{\Delta}^\top\widehat{\mathbf{W}}\right) = 0,
\end{aligned}
$$

where the last equality in the first line holds because $\mathbf{U}^\top\mathbf{\Delta}_\mathbf{U} = \mathbf{0}$ (see (E.28)) and $\left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\right\|_F^2 = \left\|\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F^2 = \sigma_r^2(\mathbf{U})$ (see (E.27)), $\widehat{\mathbf{W}}^\top\mathbf{W} = \mathbf{0}$ in the third line holds since $\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V} = \mathbf{0}$, and $\widehat{\mathbf{W}}^\top\mathbf{\Delta} = \mathbf{0}$ in the fourth and last lines holds because

$$
\widehat{\mathbf{W}}^\top\mathbf{\Delta} = \mathbf{U}^\top\mathbf{\Delta}_\mathbf{U} - \mathbf{V}^\top\mathbf{\Delta}_\mathbf{V} = \mathbf{0}.
$$

Now plugging these terms into (E.3) yields

$$
\begin{aligned}
[\nabla^2 g(\mathbf{W})](\mathbf{\Delta},\mathbf{\Delta}) &= \|\mathbf{\Delta}_{\mathbf{U}}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_{\mathbf{V}}^\top\|_F^2 + 2\langle \mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star, \mathbf{\Delta}_{\mathbf{U}}\mathbf{\Delta}_{\mathbf{V}}^\top\rangle \\
&\quad + \mu(\langle \widehat{\mathbf{W}}^\top\mathbf{W}, \widehat{\mathbf{\Delta}}^\top\mathbf{\Delta}\rangle + \langle \widehat{\mathbf{W}}\widehat{\mathbf{\Delta}}^\top, \mathbf{\Delta}\mathbf{W}^\top\rangle + \langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \mathbf{\Delta}\mathbf{\Delta}^\top\rangle) \\
&= -2(\sigma_i(\mathbf{X}^\star) - \sigma_r^2(\mathbf{U})).
\end{aligned}
$$

The proof of the strict saddle property is completed by noting that

$$
\|\mathbf{\Delta}\|_F^2 = \|\mathbf{\Delta}_{\mathbf{U}}\|_F^2 + \|\mathbf{\Delta}_{\mathbf{V}}\|_F^2 = 2,
$$

which further implies

$$
\lambda_{\min}\left(\nabla^2 g(\mathbf{W})\right) \le -2\frac{\sigma_i(\mathbf{X}^\star) - \sigma_r^2(\mathbf{U})}{\|\mathbf{\Delta}\|_F^2} \le -\left(\sigma_r(\mathbf{X}^\star) - \sigma_{r+1}(\mathbf{X}^\star)\right),
$$

where the last inequality holds because of (E.27) and because $i \le r$. □

E.11 Proof of Theorem E.11.1 (robust strict saddle for $g(\mathbf{W})$)

Theorem E.11.1. *Let $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3', \mathcal{R}_3'', \mathcal{R}_3'''$ be the regions as defined in Theorem 6.3.1. Let $g(\mathbf{W})$ be defined as in* (E.1) *with $\mu = \frac{1}{2}$. Then $g(\mathbf{W})$ has the following robust strict saddle property:*

1. *For any $\mathbf{W} \in \mathcal{R}_1$, $g(\mathbf{W})$ satisfies local regularity condition:*

$$
\langle \nabla g(\mathbf{W}), \mathbf{W} - \mathbf{W}^\star\mathbf{R}\rangle \ge \frac{1}{32}\sigma_r(\mathbf{X}^\star)\,\mathrm{dist}^2(\mathbf{W},\mathbf{W}^\star) + \frac{1}{48\|\mathbf{X}^\star\|}\|\nabla g(\mathbf{W})\|_F^2, \tag{E.12}
$$

   where $\mathrm{dist}(\mathbf{W},\mathbf{W}^\star)$ *and* $\mathbf{R}$ *are defined in* (6.12) *and* (6.13)*, respectively.*

2. *For any $\mathbf{W} \in \mathcal{R}_2$, $g(\mathbf{W})$ has a directional negative curvature:*

$$
\lambda_{\min}\left(\nabla^2 g(\mathbf{W})\right) \le -\frac{1}{4}\sigma_r(\mathbf{X}^\star). \tag{E.13}
$$

3. *For any $\mathbf{W} \in \mathcal{R}_3 = \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$, $g(\mathbf{W})$ has large gradient descent:*

$$
\|\nabla g(\mathbf{W})\|_F \ge \frac{1}{10}\sigma_r^{3/2}(\mathbf{X}^\star), \quad \forall\, \mathbf{W} \in \mathcal{R}_3'; \tag{E.14}
$$

$$
\|\nabla g(\mathbf{W})\|_F > \frac{39}{800}\|\mathbf{W}\|^3, \quad \forall\, \mathbf{W} \in \mathcal{R}_3''; \tag{E.15}
$$

$$
\langle \nabla g(\mathbf{W}), \mathbf{W}\rangle > \frac{1}{20}\left\|\mathbf{W}\mathbf{W}^\top\right\|_F^2, \quad \forall\, \mathbf{W} \in \mathcal{R}_3'''. \tag{E.16}
$$

Proof. We first establish the following useful results.

Lemma E.11.1. *For any two PSD matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, we have*

$$\sigma_n(\mathbf{A})\,\mathrm{tr}(\mathbf{B}) \le \mathrm{tr}\,(\mathbf{AB}) \le \|\mathbf{A}\|\,\mathrm{tr}(\mathbf{B}).$$

Proof of Lemma E.11.1. Let $\mathbf{A} = \mathbf{\Phi}_1\mathbf{\Lambda}_1\mathbf{\Phi}_1^\top$ and $\mathbf{B} = \mathbf{\Phi}_2\mathbf{\Lambda}_2\mathbf{\Phi}_2^\top$ be the eigendecompositions of $\mathbf{A}$ and $\mathbf{B}$, respectively. Here $\mathbf{\Lambda}_1$ ($\mathbf{\Lambda}_2$) is a diagonal matrix with the eigenvalues of $\mathbf{A}$ ($\mathbf{B}$) along its diagonal. We first rewrite $\mathrm{tr}\,(\mathbf{AB})$ as

$$\mathrm{tr}\,(\mathbf{AB}) = \mathrm{tr}\left(\mathbf{\Lambda}_1\mathbf{\Phi}_1^\top\mathbf{\Phi}_2\mathbf{\Lambda}_2\mathbf{\Phi}_2^\top\mathbf{\Phi}_1\right).$$

Noting that $\mathbf{\Lambda}_1$ is a diagonal matrix, we have

$$\begin{aligned}\mathrm{tr}\left(\mathbf{\Lambda}_1\mathbf{\Phi}_1^\top\mathbf{\Phi}_2\mathbf{\Lambda}_2\mathbf{\Phi}_2^\top\mathbf{\Phi}_1\right) &\ge \min_i \mathbf{\Lambda}_1[i,i]\cdot\mathrm{tr}\left(\mathbf{\Phi}_1^\top\mathbf{\Phi}_2\mathbf{\Lambda}_2\mathbf{\Phi}_2^\top\mathbf{\Phi}_1\right)\\ &= \sigma_n(\mathbf{A})\,\mathrm{tr}(\mathbf{B}).\end{aligned}$$

The other direction follows similarly. □

Corollary E.11.1. *For any two matrices $\mathbf{A} \in \mathbb{R}^{r\times r}$ and $\mathbf{B} \in \mathbb{R}^{n\times r}$, we have*

$$\sigma_r(\mathbf{A})\|\mathbf{B}\|_F \le \|\mathbf{AB}\|_F \le \|\mathbf{A}\|\|\mathbf{B}\|_F.$$

We provide one more result before proceeding to prove the main theorem.

Lemma E.11.2. *Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times r}$ such that $\mathbf{A}^\top\mathbf{B} = \mathbf{B}^\top\mathbf{A} \succeq \mathbf{0}$ is PSD. If $\|\mathbf{A}-\mathbf{B}\| \le \frac{\sqrt{2}}{2}\sigma_r(\mathbf{B})$, we have*

$$\underbrace{\left\langle\left(\mathbf{AA}^\top - \mathbf{BB}^\top\right)\mathbf{A}, \mathbf{A}-\mathbf{B}\right\rangle}_{(\aleph_1)} \ge \frac{1}{16}\Big(\underbrace{\mathrm{tr}((\mathbf{A}-\mathbf{B})^\top(\mathbf{A}-\mathbf{B})\mathbf{B}^\top\mathbf{B})}_{(\aleph_2)} + \underbrace{\|\mathbf{AA}^\top - \mathbf{BB}^\top\|_F^2}_{(\aleph_3)}\Big). \tag{E.29}$$

Proof. Denote $\mathbf{E} = \mathbf{A}-\mathbf{B}$. We first rewrite the terms $(\aleph_1)$, $(\aleph_2)$ and $(\aleph_3)$ as follows

$$\begin{aligned}(\aleph_1) &= \mathrm{tr}\left(\left(\mathbf{E}^\top\mathbf{E}\right)^2 + 3\mathbf{E}^\top\mathbf{E}\mathbf{E}^\top\mathbf{B} + \left(\mathbf{E}^\top\mathbf{B}\right)^2 + \mathbf{E}^\top\mathbf{E}\mathbf{B}^\top\mathbf{B}\right),\\ (\aleph_2) &= \mathrm{tr}\left(\mathbf{E}^\top\mathbf{E}\mathbf{B}^\top\mathbf{B}\right),\\ (\aleph_3) &= \mathrm{tr}\left(\left(\mathbf{E}^\top\mathbf{E}\right)^2 + 4\mathbf{E}^\top\mathbf{E}\mathbf{E}^\top\mathbf{B} + 2\left(\mathbf{E}^\top\mathbf{B}\right)^2 + 2\mathbf{E}^\top\mathbf{E}\mathbf{B}^\top\mathbf{B}\right),\end{aligned}$$

where $\mathbf{E}^\top\mathbf{B} = \mathbf{A}^\top\mathbf{B} - \mathbf{B}^\top\mathbf{B} = \mathbf{B}^\top\mathbf{E}$. Now we have

$$
\begin{aligned}
&(\aleph_1) - \frac{1}{16}(\aleph_2) - \frac{1}{16}(\aleph_3) \\
&= \operatorname{tr}\left(\frac{15}{16}\left(\mathbf{E}^\top\mathbf{E}\right)^2 + \frac{11}{4}\mathbf{E}^\top\mathbf{E}\mathbf{E}^\top\mathbf{B} + \frac{7}{8}\left(\mathbf{E}^\top\mathbf{B}\right)^2 + \frac{13}{16}\mathbf{E}^\top\mathbf{E}\mathbf{B}^\top\mathbf{B}\right) \\
&= \left\|\sqrt{\frac{121}{56}}\mathbf{E}^\top\mathbf{E} + \sqrt{\frac{7}{8}}\mathbf{E}^\top\mathbf{B}\right\|_F^2 + \operatorname{tr}\left(\frac{13}{16}\mathbf{E}^\top\mathbf{E}\mathbf{B}^\top\mathbf{B} - \frac{137}{112}\mathbf{E}^\top\mathbf{E}\mathbf{E}^\top\mathbf{E}\right) \\
&\geq \operatorname{tr}\left(\frac{13}{16}\mathbf{E}^\top\mathbf{E}\sigma_r^2(\mathbf{B}) - \frac{137}{112}\mathbf{E}^\top\mathbf{E}\|\mathbf{E}\|^2\right) \\
&\geq \operatorname{tr}\left(\left(\frac{13}{16} - \frac{137}{112}\frac{1}{2}\right)\sigma_r^2(\mathbf{B})\mathbf{E}^\top\mathbf{E}\right) \\
&\geq 0,
\end{aligned}
$$

where the third line follows from Lemma E.11.1 and the fourth line holds because by assumption $\|\mathbf{E}\| \leq \frac{\sqrt{2}}{2}\sigma_r(\mathbf{B})$.

□

Now we turn to prove the main results. Recall that $\mu = \frac{1}{2}$ throughout the proof.

E.11.1 Regularity condition for the region $\mathcal{R}_1$

It follows from Lemma 6.3.2 that $\mathbf{W}^\top\mathbf{W}^\star\mathbf{R} = \mathbf{R}^\top\mathbf{W}^{\star\top}\mathbf{W}$ is PSD, where $\mathbf{R} = \arg\min_{R'\in\mathbb{O}_r}\|\mathbf{W} - \mathbf{W}^\star\mathbf{R}'\|_F^2$. We first perform the change of variable $\mathbf{W}^\star\mathbf{R} \to \mathbf{W}^\star$ to avoid $\mathbf{R}$ in the following equations. With this change of variable we have instead $\mathbf{W}^\top\mathbf{W}^\star = \mathbf{W}^{\star\top}\mathbf{W}$ is PSD. We now rewrite the gradient $\nabla g(\mathbf{W})$ as follows:

$$
\begin{aligned}
\nabla g(\mathbf{W}) &= \begin{bmatrix} \mathbf{0} & \mathbf{U}\mathbf{V}^\top - \mathbf{U}^\star\mathbf{V}^{\star\top} \\ \mathbf{V}\mathbf{U}^\top - \mathbf{V}^\star\mathbf{U}^{\star\top} & \mathbf{0} \end{bmatrix}\mathbf{W} + \mu\widehat{\mathbf{W}}(\widehat{\mathbf{W}}^\top\mathbf{W}) \\
&= \frac{1}{2}\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W} + \frac{1}{2}\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W} + (\mu - \frac{1}{2})\widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top\mathbf{W} \\
&= \frac{1}{2}\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W} + \frac{1}{2}\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}.
\end{aligned} \tag{E.30}
$$

Plugging this into the left hand side of (E.12) gives

$$
\begin{aligned}
\langle\nabla g(\mathbf{W}), \mathbf{W} - \mathbf{W}^\star\rangle &= \frac{1}{2}\left\langle\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W}, \mathbf{W} - \mathbf{W}^\star\right\rangle + \frac{1}{2}\left\langle\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}, \mathbf{W} - \mathbf{W}^\star\right\rangle \\
&= \frac{1}{2}\left\langle\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W}, \mathbf{W} - \mathbf{W}^\star\right\rangle + \frac{1}{2}\left\langle\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \mathbf{W}\mathbf{W}^\top\right\rangle
\end{aligned} \tag{E.31}
$$

where the last line follows from the fact that $\mathbf{W}^{\star\top}\widehat{\mathbf{W}}^{\star}=\mathbf{0}$. We first show the first term in the right hand side of the above equation is sufficiently large

$$\begin{aligned}\left\langle\left(\mathbf{W}\mathbf{W}^{\top}-\mathbf{W}^{\star}\mathbf{W}^{\star\top}\right)\mathbf{W},\mathbf{W}-\mathbf{W}^{\star}\right\rangle &\geq \frac{1}{16}\operatorname{tr}\left((\mathbf{W}-\mathbf{W}^{\star})^{\top}(\mathbf{W}-\mathbf{W}^{\star})\mathbf{W}^{\star\top}\mathbf{W}^{\star}\right)\\ &\quad+\frac{1}{16}\left\|\mathbf{W}\mathbf{W}^{\top}-\mathbf{W}^{\star}\mathbf{W}^{\star\top}\right\|_F^2\\ &\geq \frac{1}{16}\sigma_r(\mathbf{W}^{\star\top}\mathbf{W}^{\star})\left\|\mathbf{W}-\mathbf{W}^{\star}\right\|_F^2\\ &\quad+\frac{1}{16}\left\|\mathbf{W}\mathbf{W}^{\top}-\mathbf{W}^{\star}\mathbf{W}^{\star\top}\right\|_F^2\\ &=\frac{1}{8}\sigma_r(\mathbf{X}^{\star})\left\|\mathbf{W}-\mathbf{W}^{\star}\right\|_F^2+\frac{1}{16}\left\|\mathbf{W}\mathbf{W}^{\top}-\mathbf{W}^{\star}\mathbf{W}^{\star\top}\right\|_F^2,\end{aligned}\tag{E.32}$$

where the first inequality follows from Lemma E.11.2 since $\mathbf{W}^{\top}\mathbf{W}^{\star}=\mathbf{W}^{\star\top}\mathbf{W}$ is PSD and $\|\mathbf{W}-\mathbf{W}^{\star}\|\leq\sigma_r^{1/2}(\mathbf{X}^{\star})=\frac{\sqrt{2}}{2}\sigma_r(\mathbf{W}^{\star})$, the second inequality follows from Lemma E.11.1, and the last line holds because

$$\sigma_r\left(\widehat{\mathbf{W}}^{\star\top}\widehat{\mathbf{W}}^{\star}\right)=\sigma_r\left(\widehat{\mathbf{U}}^{\star\top}\widehat{\mathbf{U}}^{\star}+\widehat{\mathbf{V}}^{\star\top}\widehat{\mathbf{V}}^{\star}\right)=2\sigma_r\left(\mathbf{\Sigma}\right)=2\sigma_r\left(\mathbf{X}^{\star}\right).$$

We then show the second term in the right hand side of (E.31) is lower bounded by

$$\begin{aligned}\left\langle\widehat{\mathbf{W}}^{\star}\widehat{\mathbf{W}}^{\star\top},\mathbf{W}\mathbf{W}^{\top}\right\rangle&=\frac{1}{2\left\|\mathbf{X}^{\star}\right\|}\left\|\widehat{\mathbf{W}}^{\star\top}\widehat{\mathbf{W}}^{\star}\right\|\operatorname{tr}\left(\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\mathbf{W}^{\top}\widehat{\mathbf{W}}^{\star}\right)\\ &\geq\frac{1}{2\left\|\mathbf{X}^{\star}\right\|}\operatorname{tr}\left(\widehat{\mathbf{W}}^{\star\top}\widehat{\mathbf{W}}^{\star}\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\mathbf{W}^{\top}\widehat{\mathbf{W}}^{\star}\right)\\ &=\frac{1}{2\left\|\mathbf{X}^{\star}\right\|}\left\|\widehat{\mathbf{W}}^{\star}\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2\end{aligned}\tag{E.33}$$

where the first line holds because $\left\|\widehat{\mathbf{W}}^{\star\top}\widehat{\mathbf{W}}^{\star}\right\|=\left\|\widehat{\mathbf{U}}^{\star\top}\widehat{\mathbf{U}}^{\star}+\widehat{\mathbf{V}}^{\star\top}\widehat{\mathbf{V}}^{\star}\right\|=2\left\|\mathbf{\Sigma}\right\|=2\left\|\mathbf{X}^{\star}\right\|$, and the inequality follows from Lemma E.11.1.

On the other hand, we attempt to control the gradient of $g(\mathbf{W})$. To that end, it follows from (E.30) that

$$\begin{aligned}\left\|\nabla g(\mathbf{W})\right\|_F^2&=\frac{1}{4}\left\|\left(\mathbf{W}\mathbf{W}^{\top}-\mathbf{W}^{\star}\mathbf{W}^{\star\top}\right)\mathbf{W}+\widehat{\mathbf{W}}^{\star}\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2\\ &\leq\frac{12}{47}\left\|\left(\mathbf{W}\mathbf{W}^{\top}-\mathbf{W}^{\star}\mathbf{W}^{\star\top}\right)\mathbf{W}\right\|_F^2+12\left\|\widehat{\mathbf{W}}^{\star}\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2\\ &\leq\frac{12}{47}\|\mathbf{W}\|^2\left\|\mathbf{W}\mathbf{W}^{\top}-\mathbf{W}^{\star}\mathbf{W}^{\star\top}\right\|_F^2+12\left\|\widehat{\mathbf{W}}^{\star}\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2,\end{aligned}\tag{E.34}$$

where the first inequality holds since $(a+b)^2\leq\frac{1+\epsilon}{\epsilon}a^2+(1+\epsilon)b^2$ for any $\epsilon>0$.

Combining (E.31)- (E.34), we can conclude the proof of (E.12) as long as we can show the following inequality:

$$\frac{1}{8}\left\|\mathbf{W}\mathbf{W}^{\top}-\mathbf{W}^{\star}\mathbf{W}^{\star\top}\right\|_F^2\geq\frac{1}{47}\frac{\|\mathbf{W}\|^2}{\left\|\mathbf{X}^{\star}\right\|}\left\|\mathbf{W}\mathbf{W}^{\top}-\mathbf{W}^{\star}\mathbf{W}^{\star\top}\right\|_F^2.$$

To that end, we upper bound $\|\mathbf{W}\|$ as follows:

$$\begin{aligned}\|\mathbf{W}\| &\le \|\mathbf{W}^\star\| + \|\mathbf{W}-\mathbf{W}^\star\| \\ &\le \sqrt{2}\sigma_1^{1/2}(\mathbf{X}^\star) + \|\mathbf{W}-\mathbf{W}^\star\|_F \\ &\le (\sqrt{2}+1)\sigma_1^{1/2}(\mathbf{X}^\star)\end{aligned}$$

since $\|\mathbf{W}^\star\| = \sqrt{2}\sigma_1^{(1/2)}(\mathbf{X}^\star)$ and $\mathrm{dist}(\mathbf{W},\mathbf{W}^\star) \le \sigma_r^{(1/2)}(\mathbf{X}^\star)$. This completes the proof of (E.12).

E.11.2 Negative curvature for the region $\mathcal{R}_2$

To show (E.13), we utilize a strategy similar to that used in Appendix E.8 for proving the strict saddle property of $g(\mathbf{W})$ by constructing a direction $\boldsymbol{\Delta}$ such that the Hessian evaluated at $\mathbf{W}$ along this direction is negative. For this purpose, denote

$$\mathbf{Q} = \begin{bmatrix}\boldsymbol{\Phi}/\sqrt{2} \\ \boldsymbol{\Psi}/\sqrt{2}\end{bmatrix}, \tag{E.35}$$

where we recall that $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ consist of the left and right singular vectors of $\mathbf{X}^\star$, respectively. The optimal solution $\mathbf{W}^\star$ has a compact SVD $\mathbf{W}^\star = \mathbf{Q}(\sqrt{2}\boldsymbol{\Sigma}^{1/2})\mathbf{R}$. For notational convenience, we denote $\overline{\boldsymbol{\Sigma}} = 2\boldsymbol{\Sigma}$, where $\overline{\boldsymbol{\Sigma}}$ is a diagonal matrix whose diagonal entries in the upper left corner are $\overline{\sigma}_1,\dots,\overline{\sigma}_r$.

For any $\mathbf{W}$, we can always divide it into two parts, the projections onto the column spaces of $\mathbf{Q}$ and its orthogonal complement, respectively. Equivalently, we can write

$$\mathbf{W} = \mathbf{Q}\overline{\boldsymbol{\Lambda}}^{1/2}\mathbf{R} + \mathbf{E}, \tag{E.36}$$

where $\mathbf{Q}\overline{\boldsymbol{\Lambda}}^{1/2}\mathbf{R}$ is a compact SVD form representing the projection of $\mathbf{W}$ onto the column space of $\mathbf{Q}$, and $\mathbf{E}^\top\mathbf{Q} = \mathbf{0}$ (i.e., $\mathbf{E}$ is orthogonal to $\mathbf{Q}$). Here $\mathbf{R} \in \mathcal{O}_r$ and $\overline{\boldsymbol{\Lambda}}$ is a diagonal matrix whose diagonal entries in the upper left corner are $\overline{\lambda}_1,\dots,\overline{\lambda}_r$, but the diagonal entries are not necessarily placed either in decreasing or increasing order. In order to characterize the neighborhood near all strict saddles $\mathcal{C}\setminus\mathcal{X}$, we consider $\mathbf{W}$ such that $\sigma_r(\mathbf{W}) \le \sqrt{\frac{3}{8}}\sigma_r^{1/2}(\mathbf{X}^\star)$. Let $k := \arg\min_i \overline{\lambda}_i$ denote the location of the smallest diagonal entry in $\overline{\boldsymbol{\Lambda}}$. It is clear that

$$\overline{\lambda}_k \le \sigma_r^2(\mathbf{W}) \le \frac{3}{8}\sigma_r(\mathbf{X}^\star). \tag{E.37}$$

Let $\boldsymbol{\alpha} \in \mathbb{R}^r$ be the eigenvector associated with the smallest eigenvalue of $\mathbf{W}^\top\mathbf{W}$.

Recall that $\mu = \frac{1}{2}$. We show that the function $g(\mathbf{W})$ at $\mathbf{W}$ has directional negative curvature along the direction

$$\boldsymbol{\Delta} = \mathbf{q}_k\boldsymbol{\alpha}^\top. \tag{E.38}$$

We repeat the Hessian evaluated at $\mathbf{W}$ for $\boldsymbol{\Delta}$ as follows

$$
\begin{aligned}
&[\nabla^2 g(\mathbf{W})](\boldsymbol{\Delta},\boldsymbol{\Delta}) \\
&= \underbrace{\left\|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top + \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\|_F^2}_{\Pi_1} + 2\underbrace{\left\langle \mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star, \boldsymbol{\Delta}_{\mathbf{U}}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\rangle}_{\Pi_2} + \frac{1}{2}\underbrace{\left\langle \widehat{\boldsymbol{\Delta}}\widehat{\mathbf{W}}^\top, \boldsymbol{\Delta}\mathbf{W}^\top\right\rangle}_{\Pi_3} + \frac{1}{2}\underbrace{\left\langle \widehat{\mathbf{W}}\widehat{\boldsymbol{\Delta}}^\top, \boldsymbol{\Delta}\mathbf{W}^\top\right\rangle}_{\Pi_4} + \frac{1}{2}\underbrace{\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right\rangle}_{\Pi_5}
\end{aligned}
$$

The remaining part is to bound the five terms.

Bounding terms Π_1, Π_3 and Π_4. We first rewrite these three terms:

$$
\begin{aligned}
\Pi_1 &= \|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top\|_F^2 + \|\mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top\|_F^2 + 2\left\langle \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top, \boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top\right\rangle, \\
\Pi_3 &= \left\langle \widehat{\boldsymbol{\Delta}}\widehat{\mathbf{W}}^\top, \boldsymbol{\Delta}\mathbf{W}^\top\right\rangle = \|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{U}^\top\|_F^2 + \|\boldsymbol{\Delta}_{\mathbf{V}}\mathbf{V}^\top\|_F^2 - \|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top\|_F^2 - \|\boldsymbol{\Delta}_{\mathbf{V}}\mathbf{U}^\top\|_F^2, \\
\Pi_4 &= \left\langle \mathbf{U}\boldsymbol{\Delta}_{\mathbf{U}}^\top, \boldsymbol{\Delta}_{\mathbf{U}}\mathbf{U}^\top\right\rangle + \left\langle \mathbf{V}\boldsymbol{\Delta}_{\mathbf{V}}^\top, \boldsymbol{\Delta}_{\mathbf{V}}\mathbf{V}^\top\right\rangle - 2\left\langle \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top, \boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top\right\rangle \\
&\leq \|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{U}^\top\|_F^2 + \|\boldsymbol{\Delta}_{\mathbf{V}}\mathbf{V}^\top\|_F^2 - 2\left\langle \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top, \boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top\right\rangle,
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\Pi_1 + \frac{1}{2}\Pi_3 + \frac{1}{2}\Pi_4 \\
&\leq \|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top\|_F^2 + \|\mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top\|_F^2 + \|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{U}^\top\|_F^2 + \|\boldsymbol{\Delta}_{\mathbf{V}}\mathbf{V}^\top\|_F^2 - \frac{1}{2}\|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top\|_F^2 - \frac{1}{2}\|\boldsymbol{\Delta}_{\mathbf{V}}\mathbf{U}^\top\|_F^2 + \left\langle \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top, \boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top\right\rangle \\
&= \|\mathbf{W}\boldsymbol{\Delta}^\top\|_F^2 - \frac{1}{2}\left\|\boldsymbol{\Delta}_{\mathbf{U}}\mathbf{V}^\top - \mathbf{U}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\|_F^2 \\
&\leq \|\mathbf{W}\boldsymbol{\Delta}^\top\|_F^2.
\end{aligned} \tag{E.39}
$$

Noting that $\boldsymbol{\Delta}^\top\boldsymbol{\Delta} = \boldsymbol{\alpha}\mathbf{q}_k^\top\mathbf{q}_k\boldsymbol{\alpha}^\top = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$, we now compute $\|\mathbf{W}\boldsymbol{\Delta}^\top\|_F^2$ as

$$
\begin{aligned}
\|\mathbf{W}\boldsymbol{\Delta}^\top\|_F^2 &= \mathrm{tr}\left(\mathbf{W}^\top\mathbf{W}\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\right) = \mathrm{tr}\left(\mathbf{W}^\top\mathbf{W}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\right) \\
&= \sigma_r^2(\mathbf{W}).
\end{aligned}
$$

Plugging this into (E.39) gives

$$
\Pi_1 + \frac{1}{2}\Pi_3 + \frac{1}{2}\Pi_4 \leq \sigma_r^2(\mathbf{W}). \tag{E.40}
$$

Bounding terms Π_2 and Π_5. To obtain an upper bound for the term Π_2, we first rewrite it as follows

$$
\begin{aligned}
\Pi_2 &= \left\langle \mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star, \boldsymbol{\Delta}_{\mathbf{U}}\boldsymbol{\Delta}_{\mathbf{V}}^\top\right\rangle \\
&= \frac{1}{2}\left\langle \begin{bmatrix} \mathbf{0} & \mathbf{U}\mathbf{V}^\top - \mathbf{U}^\star\mathbf{V}^{\star\top} \\ \mathbf{V}\mathbf{U}^\top - \mathbf{V}^\star\mathbf{U}^{\star\top} & \mathbf{0} \end{bmatrix}, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right\rangle \\
&= \frac{1}{4}\left\langle \mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right\rangle - \frac{1}{4}\left\langle \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right\rangle + \frac{1}{4}\left\langle \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right\rangle.
\end{aligned}
$$

We then have

$$2\Pi_2 + \frac{1}{2}\Pi_5 = \frac{1}{2}\left\langle \mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top \right\rangle + \frac{1}{2}\left\langle \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top \right\rangle. \tag{E.41}$$

To bound these two terms in the above equation, we note that

$$\boldsymbol{\Delta}\boldsymbol{\Delta}^\top = \sum_{i=1}^{r} \alpha_i^2 \mathbf{q}_k\mathbf{q}_k^\top = \mathbf{q}_k\mathbf{q}_k^\top = \frac{1}{2}\begin{bmatrix} \boldsymbol{\phi}_k\boldsymbol{\phi}_k^\top & \boldsymbol{\phi}_k\boldsymbol{\psi}_k^\top \\ \boldsymbol{\psi}_k\boldsymbol{\phi}_k^\top & \boldsymbol{\psi}_k\boldsymbol{\psi}_k^\top \end{bmatrix}.$$

Then we have

$$\left\langle \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top \right\rangle = \frac{1}{2}\left\langle \begin{bmatrix} \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top & -\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^\top \\ -\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top & \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^\top \end{bmatrix}, \begin{bmatrix} \boldsymbol{\phi}_k\boldsymbol{\phi}_k^\top & \boldsymbol{\phi}_k\boldsymbol{\psi}_k^\top \\ \boldsymbol{\psi}_k\boldsymbol{\phi}_k^\top & \boldsymbol{\psi}_k\boldsymbol{\psi}_k^\top \end{bmatrix} \right\rangle = 0,$$

and

$$\begin{aligned} \left\langle \mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top \right\rangle &= \left\langle \mathbf{Q}\overline{\boldsymbol{\Lambda}}\mathbf{Q}^\top - 2\mathbf{Q}\overline{\boldsymbol{\Lambda}}^{1/2}\mathbf{R}\mathbf{E}^\top + \mathbf{E}\mathbf{E}^\top - \mathbf{Q}\overline{\boldsymbol{\Sigma}}\mathbf{Q}^\top, \mathbf{q}_k\mathbf{q}_k^\top \right\rangle \\ &= \overline{\lambda}_k - \overline{\sigma}_k \end{aligned}$$

where the last utilizes the fact that $\mathbf{E}^\top\mathbf{q}_k = \mathbf{0}$ since $\mathbf{E}$ is orthogonal to $\mathbf{Q}$.

Plugging these into (E.41) gives

$$2\Pi_2 + \frac{1}{2}\Pi_5 = \frac{1}{2}(\overline{\lambda}_k - \overline{\sigma}_k). \tag{E.42}$$

Merging together. Putting (E.40) and (E.42) together yields

$$\begin{aligned} [\nabla^2 g(\mathbf{W})](\boldsymbol{\Delta}, \boldsymbol{\Delta}) &= \Pi_1 + \frac{1}{2}\Pi_3 + \frac{1}{2}\Pi_4 + 2\Pi_2 + \frac{1}{2}\Pi_5 \\ &\le \sigma_r^2(\mathbf{W}) + \frac{1}{2}(\overline{\lambda}_k - \overline{\sigma}_k) \\ &\le \frac{1}{2}\sigma_r(\mathbf{X}^\star) + \frac{1}{2}(\frac{1}{2}\sigma_r(\mathbf{X}^\star) - 2\sigma_r(\mathbf{X}^\star)) \\ &\le -\frac{1}{4}\sigma_r(\mathbf{X}^\star), \end{aligned}$$

where the third line follows because by assumption $\sigma_r(\mathbf{W}) \le \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\mathbf{X}^\star)$, by construction $\overline{\lambda}_k \le \frac{1}{2}\sigma_r(\mathbf{X}^\star)$ (see (E.37)), and $\overline{\sigma}_k \ge \overline{\sigma}_r = 2\sigma_r(\mathbf{X}^\star)$. This completes the proof of (E.13).

E.11.3 Large gradient for the region $\mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$

In order to show that $g(\mathbf{W})$ has a large gradient in the three regions $\mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$, we first provide a lower bound for the gradient. By (E.30), we have

$$\begin{aligned}
\|\nabla g(\mathbf{W})\|_F^2 &= \frac{1}{4}\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W} + \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2 \\
&= \frac{1}{4}\left(\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W}\right\|_F^2 + \left\|\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2\right) + \frac{1}{2}\left\langle\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W}, \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\rangle \\
&= \frac{1}{4}\left(\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W}\right\|_F^2 + \left\|\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2\right) + \frac{1}{2}\left\langle\mathbf{W}\mathbf{W}^\top\mathbf{W}\mathbf{W}^\top, \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\right\rangle \\
&\geq \frac{1}{4}\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W}\right\|_F^2,
\end{aligned} \tag{E.43}$$

where the third equality follows because $\mathbf{W}^{\star\top}\widehat{\mathbf{W}}^\star = \mathbf{U}^{\star\top}\mathbf{U}^\star - \mathbf{V}^{\star\top}\mathbf{V}^\star = \mathbf{0}$ and the last line utilizes the fact that the inner product between two PSD matrices is nonnegative.

E.11.3.1 Large gradient for the region $\mathcal{R}_3'$

To show $\|\nabla g(\mathbf{W})\|_F^2$ is large for any $\mathbf{W} \in \mathcal{R}_3'$, again, for any $\mathbf{W} \in \mathbb{R}^{(n+m)\times r}$, we utilize (E.36) to write $\mathbf{W} = \mathbf{Q}\overline{\mathbf{\Lambda}}^{1/2}\mathbf{R} + \mathbf{E}$, where $\mathbf{Q}$ is defined in (E.35), $\mathbf{Q}\overline{\mathbf{\Lambda}}^{1/2}\mathbf{R}$ is a compact SVD form representing the projection of $\mathbf{W}$ onto the column space of $\mathbf{Q}$, and $\mathbf{E}^\top\mathbf{Q} = \mathbf{0}$ (i.e., $\mathbf{E}$ is orthogonal to $\mathbf{Q}$). Plugging this form of $\mathbf{W}$ into the last term of (E.43) gives

$$\begin{aligned}
\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W}\right\|_F^2 &= \left\|\mathbf{Q}\overline{\mathbf{\Lambda}}^{1/2}(\overline{\mathbf{\Lambda}} - \overline{\mathbf{\Sigma}})\mathbf{R} + \mathbf{Q}\overline{\mathbf{\Lambda}}^{1/2}\mathbf{R}\mathbf{E}\mathbf{E}^\top + \mathbf{E}\mathbf{R}^\top\overline{\mathbf{\Lambda}}\mathbf{R} + \mathbf{E}\mathbf{E}^\top\mathbf{E}\right\|_F^2 \\
&= \left\|\mathbf{Q}\overline{\mathbf{\Lambda}}^{1/2}(\overline{\mathbf{\Lambda}} - \overline{\mathbf{\Sigma}})\mathbf{R} + \mathbf{Q}\overline{\mathbf{\Lambda}}^{1/2}\mathbf{R}\mathbf{E}\mathbf{E}^\top\right\|_F^2 \\
&\quad + \left\|\mathbf{E}\mathbf{R}^\top\overline{\mathbf{\Lambda}}\mathbf{R} + \mathbf{E}\mathbf{E}^\top\mathbf{E}\right\|_F^2
\end{aligned} \tag{E.44}$$

since $\mathbf{Q}$ is orthogonal to $\mathbf{E}$. The remaining part is to show at least one of the two terms is large for any $\mathbf{W} \in \mathcal{R}_3'$ by considering the following two cases.

Case I: $\|\mathbf{E}\|_F^2 \geq \frac{4}{25}\sigma_r(\mathbf{X}^\star)$. As $\mathbf{E}$ is large, we bound the second term in (E.44):

$$\begin{aligned}
\left\|\mathbf{E}\mathbf{R}^\top\overline{\mathbf{\Lambda}}\mathbf{R} + \mathbf{E}\mathbf{E}^\top\mathbf{E}\right\|_F^2 &\geq \sigma_r^2\left(\mathbf{R}^\top\overline{\mathbf{\Lambda}}\mathbf{R} + \mathbf{E}^\top\mathbf{E}\right)\|\mathbf{E}\|_F^2 \\
&= \sigma_r^4\left(\mathbf{W}\right)\|\mathbf{E}\|_F^2 \\
&\geq (\frac{1}{2})^2\frac{4}{25}\sigma_r^3(\mathbf{X}^\star) = \frac{1}{25}\sigma_r^3(\mathbf{X}^\star),
\end{aligned} \tag{E.45}$$

where the first inequality follows from Corollary E.11.1, the first equality follows from the fact $\mathbf{W}^\top\mathbf{W} = \mathbf{R}^\top\overline{\mathbf{\Lambda}}\mathbf{R} + \mathbf{E}^\top\mathbf{E}$, and the last inequality holds because by assumption that $\sigma_r^2(\mathbf{W}) \geq \frac{1}{2}\sigma_r(\mathbf{X}^\star)$ and $\|\mathbf{E}\|_F^2 \geq \frac{4}{25}\sigma_r(\mathbf{X}^\star)$.

Case II: $\|\mathbf{E}\|_F^2 \leq \frac{4}{25}\sigma_r(\mathbf{X}^\star)$. In this case, we start by bounding the diagonal entries in $\overline{\mathbf{\Lambda}}$. First, utilizing Weyl's inequality for perturbation of singular values [271, Theorem 3.3.16] gives

$$\left|\sigma_r(\mathbf{W}) - \min_i \overline{\lambda}_i^{1/2}\right| \leq \|\mathbf{E}\|_2,$$

which implies

$$\min_i \overline{\lambda}_i^{1/2} \geq \sigma_r(\mathbf{W}) - \|\mathbf{E}\|_2 \geq \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\mathbf{X}^\star) - \frac{2}{5}\sigma_r^{1/2}(\mathbf{X}^\star), \tag{E.46}$$

where we utilize $\|\mathbf{E}\|_2 \leq \|\mathbf{E}\|_F \leq \frac{2}{5}\sigma_r^{1/2}(\mathbf{X}^\star)$. On the other hand,

$$\begin{aligned}\operatorname{dist}(\mathbf{W}, \mathbf{W}^\star) &\leq \left\|\mathbf{Q}(\overline{\mathbf{\Lambda}}^{1/2} - \overline{\mathbf{\Sigma}}^{1/2})\mathbf{R} + \mathbf{E}\right\|_F \\ &\leq \left\|\mathbf{Q}(\overline{\mathbf{\Lambda}}^{1/2} - \overline{\mathbf{\Sigma}}^{1/2})\mathbf{R}\right\|_F + \|\mathbf{E}\|_F,\end{aligned}$$

which together with the assumption that $\operatorname{dist}(\mathbf{W}, \mathbf{W}^\star) \geq \sigma_r^{1/2}(\mathbf{X}^\star)$ gives

$$\left\|\overline{\mathbf{\Lambda}}^{1/2} - \overline{\mathbf{\Sigma}}^{1/2}\right\|_F \geq \sigma_r^{1/2}(\mathbf{X}^\star) - \frac{2}{5}\sigma_r^{1/2}(\mathbf{X}^\star) = \frac{3}{5}\sigma_r^{1/2}(\mathbf{X}^\star).$$

We now bound the first term in (E.44):

$$\begin{aligned}\left\|\mathbf{Q}\overline{\mathbf{\Lambda}}^{1/2}(\overline{\mathbf{\Lambda}} - \overline{\mathbf{\Sigma}})\mathbf{R} + \mathbf{Q}\overline{\mathbf{\Lambda}}^{1/2}\mathbf{R}\mathbf{E}\mathbf{E}^\top\right\|_F &\geq \min_i \overline{\lambda}_i^{1/2}\left\|(\overline{\mathbf{\Lambda}} - \overline{\mathbf{\Sigma}})\mathbf{R} + \mathbf{R}\mathbf{E}\mathbf{E}^\top\right\|_F \\ &\geq \min_i \overline{\lambda}_i^{1/2}\left(\left\|(\overline{\mathbf{\Lambda}} - \overline{\mathbf{\Sigma}})\mathbf{R}\right\|_F - \left\|\mathbf{R}\mathbf{E}\mathbf{E}^\top\right\|_F\right) \\ &\geq (\sqrt{\frac{1}{2}} - \frac{2}{5})\left(\left(\sqrt{2} + \sqrt{\frac{1}{2}} - \frac{2}{5}\right)\frac{3}{5} - \frac{4}{25}\right)\sigma_r^{3/2}(\mathbf{X}^\star)\end{aligned} \tag{E.47}$$

where the third line holds because $\|\mathbf{E}\mathbf{E}^\top\|_F \leq \|\mathbf{E}\|_F^2 \leq \frac{4}{25}\sigma_r(\mathbf{X}^\star)$, $\min_i \overline{\lambda}_i^{1/2} \geq \left(\sqrt{\frac{1}{2}} - \frac{2}{5}\right)\sigma_r^{1/2}(\mathbf{X}^\star)$ by (E.46), and

$$\begin{aligned}\left\|\overline{\mathbf{\Lambda}} - \overline{\mathbf{\Sigma}}\right\|_F &= \sqrt{\sum_{i=1}^r \left(\overline{\sigma}_i - \overline{\lambda}_i\right)^2} \\ &= \sqrt{\sum_{i=1}^r \left(\overline{\sigma}_i^{1/2} - \overline{\lambda}_i^{1/2}\right)^2\left(\overline{\sigma}_i^{1/2} + \overline{\lambda}_i^{1/2}\right)^2} \\ &\geq \left(\overline{\sigma}_r^{1/2} + \min_i \overline{\lambda}_i^{1/2}\right)\sqrt{\sum_{i=1}^r \left(\overline{\sigma}_i^{1/2} - \overline{\lambda}_i^{1/2}\right)^2} \\ &= \left(\overline{\sigma}_r^{1/2} + \min_i \overline{\lambda}_i^{1/2}\right)\left\|\overline{\mathbf{\Lambda}}^{1/2} - \overline{\mathbf{\Sigma}}^{1/2}\right\|_F \\ &\geq \left(\sqrt{2} + \sqrt{\frac{1}{2}} - \frac{2}{5}\right)\frac{3}{5}\sigma_r(\mathbf{X}^\star).\end{aligned}$$

Combining (E.43) with (E.44), (E.45) and (E.47) gives

$$\|\nabla g(\mathbf{W})\|_F \geq \frac{1}{10}\sigma_r^{3/2}(\mathbf{X}^\star).$$

This completes the proof of (E.14).

E.11.3.2 Large gradient for the region $\mathcal{R}_3''$

By (E.43), we have

$$\|\nabla g(\mathbf{W})\|_F \geq \frac{1}{2}\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W}\right\|_F^2.$$

Now (E.15) follows directly from the fact $\|\mathbf{W}\| > \frac{20}{19}\|\mathbf{W}^\star\|$ and the following result.

Lemma E.11.3. *For any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times r}$ with $\|\mathbf{A}\| \geq \alpha\|\mathbf{B}\|$ and $\alpha > 1$, we have*

$$\left\|\left(\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\right)\mathbf{A}\right\|_F \geq (1 - \frac{1}{\alpha^2})\|\mathbf{A}\|^3.$$

Proof. Let $\mathbf{A} = \mathbf{\Phi}_1\mathbf{\Lambda}_1\mathbf{R}_1^\top$ and $\mathbf{B} = \mathbf{\Phi}_2\mathbf{\Lambda}_2\mathbf{R}_2^\top$ be the SVDs of $\mathbf{A}$ and $\mathbf{B}$, respectively. Then

$$\begin{aligned}
\left\|\left(\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\right)\mathbf{A}\right\|_F &= \left\|\mathbf{\Phi}_1\mathbf{\Lambda}_1^3 - \mathbf{\Phi}_2\mathbf{\Lambda}_2^2\mathbf{\Phi}_2^\top\mathbf{\Phi}_1\mathbf{\Lambda}_1\right\|_F \\
&\geq \left\|\mathbf{\Lambda}_1^3 - \mathbf{\Phi}_1^\top\mathbf{\Phi}_2\mathbf{\Lambda}_2^2\mathbf{\Phi}_2^\top\mathbf{\Phi}_1\mathbf{\Lambda}_1\right\|_F \\
&\geq \left\|\mathbf{\Lambda}_1^3 - \mathbf{\Lambda}_2^2\mathbf{\Lambda}_1\right\|_F \\
&\geq (1 - \frac{1}{\alpha^2})\|\mathbf{A}\|^3.
\end{aligned}$$

□

E.11.3.3 Large gradient for the region $\mathcal{R}_3'''$

By (E.30), we have

$$\begin{aligned}
\langle\nabla g(\mathbf{W}), \mathbf{W}\rangle &= \left\langle\frac{1}{2}\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W} + \frac{1}{2}\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}, \mathbf{W}\right\rangle \\
&\geq \frac{1}{2}\left\langle\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W}, \mathbf{W}\right\rangle \\
&\geq \frac{1}{2}\left(\left\|\mathbf{W}\mathbf{W}^\top\right\|_F^2 - \left\|\mathbf{W}\mathbf{W}^\top\right\|_F\left\|\mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F\right) \\
&> \frac{1}{20}\left\|\mathbf{W}\mathbf{W}^\top\right\|_F^2
\end{aligned} \tag{E.48}$$

where the last line holds because $\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F < \frac{9}{10}\|\mathbf{W}\mathbf{W}^\top\|_F$.

□

E.12 Proof of Theorem 6.3.1 (robust strict saddle for $G(\mathbf{W})$)

Theorem E.12.1 (Theorem 6.3.1)**.** *Define the following regions*

$$
\begin{aligned}
\mathcal{R}_1 :&= \left\{\mathbf{W} : \operatorname{dist}(\mathbf{W}, \mathbf{W}^\star) \le \sigma_r^{1/2}(\mathbf{X}^\star)\right\},\\
\mathcal{R}_2 :&= \left\{\mathbf{W} : \sigma_r(\mathbf{W}) \le \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\mathbf{X}^\star), \|\mathbf{W}\mathbf{W}^\top\|_F \le \frac{20}{19}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F\right\},\\
\mathcal{R}_3' :&= \left\{\mathbf{W} : \operatorname{dist}(\mathbf{W}, \mathbf{W}^\star) > \sigma_r^{1/2}(\mathbf{X}^\star), \|\mathbf{W}\| \le \frac{20}{19}\|\mathbf{W}^\star\|, \sigma_r(\mathbf{W}) > \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\mathbf{X}^\star), \|\mathbf{W}\mathbf{W}^\top\|_F \le \frac{20}{19}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F\right\},\\
\mathcal{R}_3'' :&= \left\{\mathbf{W} : \|\mathbf{W}\| > \frac{20}{19}\|\mathbf{W}^\star\| = \frac{20}{19}\sqrt{2}\|\mathbf{X}^\star\|^{1/2}, \|\mathbf{W}\mathbf{W}^\top\|_F \le \frac{10}{9}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F\right\},\\
\mathcal{R}_3''' :&= \left\{\mathbf{W} : \|\mathbf{W}\mathbf{W}^\top\|_F > \frac{10}{9}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F = \frac{20}{9}\|\mathbf{X}^\star\|_F\right\}.
\end{aligned}
$$

Let $G(\mathbf{W})$ be defined as in (6.9) *with $\mu = \frac{1}{2}$. Suppose $f(\mathbf{X})$ has a critical point $\mathbf{X}^\star \in \mathbb{R}^{n\times m}$ of rank r and satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (6.6) *with positive constants $a = 1 - c, b = 1 + c$ and*

$$
c \lesssim \frac{\sigma_r^{3/2}(\mathbf{X}^\star)}{\|\mathbf{X}^\star\|_F\|\mathbf{X}^\star\|^{1/2}}. \tag{6.14}
$$

Then $G(\mathbf{W})$ has the following robust strict saddle property:

1. *For any $\mathbf{W} \in \mathcal{R}_1$, $G(\mathbf{W})$ satisfies the local regularity condition:*

$$
\langle \nabla G(\mathbf{W}), \mathbf{W} - \mathbf{W}^\star\rangle \gtrsim \sigma_r(\mathbf{X}^\star)\operatorname{dist}^2(\mathbf{W}, \mathbf{W}^\star) + \frac{1}{\|\mathbf{X}^\star\|}\|\nabla G(\mathbf{W})\|_F^2, \tag{6.15}
$$

   where $\operatorname{dist}(\mathbf{W}, \mathbf{W}^\star)$ and $\mathbf{R}$ are defined in (6.12) *and* (6.13)*, respectively.*

2. *For any $\mathbf{W} \in \mathcal{R}_2$, $G(\mathbf{W})$ has a directional negative curvature, i.e.,*

$$
\lambda_{\min}\left(\nabla^2 G(\mathbf{W})\right) \lesssim -\sigma_r(\mathbf{X}^\star). \tag{6.16}
$$

3. *For any $\mathbf{W} \in \mathcal{R}_3 = \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$, $G(\mathbf{W})$ has large gradient descent:*

$$
\begin{aligned}
&\|\nabla G(\mathbf{W})\|_F \gtrsim \sigma_r^{3/2}(\mathbf{X}^\star), \quad \forall\, \mathbf{W} \in \mathcal{R}_3'; && (6.17)\\
&\|\nabla G(\mathbf{W})\|_F \gtrsim \|\mathbf{W}\|^3, \quad \forall\, \mathbf{W} \in \mathcal{R}_3''; && (6.18)\\
&\|\nabla G(\mathbf{W})\|_F \gtrsim \sigma_r(\mathbf{X}^\star)\left(\|\mathbf{W}\mathbf{W}^\top\|_F\right)^{1/2}, \quad \forall\, \mathbf{W} \in \mathcal{R}_3'''. && (6.19)
\end{aligned}
$$

Proof. Throughout the proofs, we always utilize $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ unless stated otherwise. To give a sense that the geometric result in Theorem E.11.1 for $g(\mathbf{W})$ is also possibly preserved for $G(\mathbf{W})$, we first compute the derivative of $G(\mathbf{W})$ as

$$\nabla G(\mathbf{W}) = \begin{bmatrix} \nabla f(\mathbf{U}\mathbf{V}^\top)\mathbf{V} \\ (\nabla f(\mathbf{U}\mathbf{V}^\top))^\top \mathbf{U} \end{bmatrix} + \mu \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top \mathbf{W}. \tag{E.49}$$

For any $\mathbf{\Delta} = \begin{bmatrix} \mathbf{\Delta}_\mathbf{U} \\ \mathbf{\Delta}_\mathbf{V} \end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$, algebraic calculation gives the Hessian quadrature form $[\nabla^2 G(\mathbf{W})](\mathbf{\Delta}, \mathbf{\Delta})$ as

$$\begin{aligned}[\nabla^2 G(\mathbf{W})](\mathbf{\Delta}, \mathbf{\Delta}) = [\nabla^2 f(\mathbf{U}\mathbf{V}^\top)](\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top, \mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top) \\ + 2\langle \nabla f(\mathbf{U}\mathbf{V}^\top), \mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top \rangle + [\nabla^2 \rho(\mathbf{W})](\mathbf{\Delta}, \mathbf{\Delta})\end{aligned} \tag{E.50}$$

where $[\nabla^2 \rho(\mathbf{W})](\mathbf{\Delta}, \mathbf{\Delta})$ is defined in (E.4). Thus, it is expected that $G(\mathbf{W})$, $\nabla G(\mathbf{W})$, and $\nabla^2 G(\mathbf{W})$ are close to their counterparts (i.e., $g(\mathbf{W})$, $\nabla g(\mathbf{W})$ and $\nabla^2 g(\mathbf{W})$) for the matrix factorization problem when $f(\mathbf{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (6.6).

Before moving to the main proofs, we provide several useful results regarding the deviations of the gradient and Hessian. We start with a useful characterization of the restricted strong convexity and smoothness condition.

Lemma E.12.1. *Suppose f satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (6.6) *with positive constants $a = 1 - c$ and $b = 1 + c, c \in [0, 1)$. Then any $n \times m$ matrices $\mathbf{C}, \mathbf{D}, \mathbf{H}$ with $\operatorname{rank}(\mathbf{C}), \operatorname{rank}(\mathbf{D}) \le r$ and $\operatorname{rank}(\mathbf{H}) \le 2r$, we have*

$$|\langle \nabla f(\mathbf{C}) - \nabla f(\mathbf{D}) - (\mathbf{C} - \mathbf{D}), \mathbf{H} \rangle| \le c \|\mathbf{C} - \mathbf{D}\|_F \|\mathbf{H}\|_F.$$

Proof of Lemma E.12.1. We first invoke [8, Proposition 2] which states that under Assumption 6.3.2 for any $n \times m$ matrices $\mathbf{Z}, \mathbf{D}, \mathbf{H}$ of rank at most $2r$, we have

$$\left| [\nabla^2 f(\mathbf{Z})](\mathbf{D}, \mathbf{H}) - \langle \mathbf{D}, \mathbf{H} \rangle \right| \le c \|\mathbf{D}\|_F \|\mathbf{H}\|_F. \tag{E.51}$$

Now using integral form of the mean value theorem for ∇f, we have

$$\begin{aligned}
&|\langle \nabla f(\mathbf{C}) - \nabla f(\mathbf{D}) - (\mathbf{C} - \mathbf{D}), \mathbf{H} \rangle| \\
&= \left| \int_0^1 \left[\nabla^2 f(t\mathbf{C} + (1-t)\mathbf{D}) \right] (\mathbf{C} - \mathbf{D}, \mathbf{H}) - \langle \mathbf{C} - \mathbf{D}, \mathbf{H} \rangle dt \right| \\
&\le \int_0^1 \left| \left[\nabla^2 f(t\mathbf{C} + (1-t)\mathbf{D}) \right] (\mathbf{C} - \mathbf{D}, \mathbf{H}) - \langle \mathbf{C} - \mathbf{D}, \mathbf{H} \rangle \right| dt \\
&\le \int_0^1 c \|\mathbf{C} - \mathbf{D}\|_F \|\mathbf{H}\|_F \, dt = c \|\mathbf{C} - \mathbf{D}\|_F \|\mathbf{H}\|_F.
\end{aligned}$$

where the second inequality follows from (E.51) since $t\mathbf{C} + (1-t)\mathbf{D}$, $\mathbf{C} - \mathbf{D}$, and $\mathbf{H}$ all are rank at most $2r$.

□

The following result controls the deviation of the gradient between the general low-rank optimization (6.9) and the matrix factorization problem by utilizing the $(2r, 4r)$-restricted strong convexity and smoothness condition (6.6).

Lemma E.12.2. *Suppose $f(\mathbf{X})$ has a critical point $\mathbf{X}^\star \in \mathbb{R}^{n\times m}$ of rank r and satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (6.6) *with positive constants $a = 1 - c$ and $b = 1 + c, c \in [0, 1)$. Then, we have*

$$\|\nabla G(\mathbf{W}) - \nabla g(\mathbf{W})\|_F \le c \left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F \|\mathbf{W}\|.$$

Proof of Lemma E.12.2. We bound the deviation directly:

$$\begin{aligned}
\|\nabla G(\mathbf{W}) - \nabla g(\mathbf{W})\|_F &= \max_{\|\mathbf{\Delta}\|_F=1} \langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{\Delta}\rangle \\
&= \max_{\|\mathbf{\Delta}\|_F=1} \left\langle \nabla f(\mathbf{X}), \mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\right\rangle - \left\langle \mathbf{X} - \mathbf{X}^\star, \mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\right\rangle \\
&\quad + \left\langle \nabla f(\mathbf{X}), \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\rangle - \left\langle \mathbf{X} - \mathbf{X}^\star, \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\rangle \\
&= \max_{\|\mathbf{\Delta}\|_F=1} \left\langle \nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^\star) - (\mathbf{X} - \mathbf{X}^\star), \mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\right\rangle \\
&\quad + \left\langle \nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^\star) - (\mathbf{X} - \mathbf{X}^\star), \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\rangle \\
&\le \max_{\|\mathbf{\Delta}\|_F=1} c\|\mathbf{X} - \mathbf{X}^\star\|_F \left(\left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\right\|_F + \left\|\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F\right) \\
&\le c\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\|_F \left(\|\mathbf{V}\| + \|\mathbf{U}\|\right) \\
&\le c\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\|_F \|\mathbf{W}\|,
\end{aligned}$$

where the last equality follows from Assumption 6.3.1 that $\nabla f(\mathbf{X}^\star) = \mathbf{0}$ and and the first inequality utilizes Lemma E.12.1.

□

Similarly, the next result controls the deviation of the Hessian between the matrix sensing problem and the matrix factorization problem.

Lemma E.12.3. *Suppose $f(\mathbf{X})$ has a critical point $\mathbf{X}^\star \in \mathbb{R}^{n\times m}$ of rank r and satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition* (6.6) *with positive constants $a = 1 - c$ and $b = 1 + c, c \in [0, 1)$. Then, for any $\mathbf{\Delta} = \begin{bmatrix}\mathbf{\Delta}_\mathbf{U}\\ \mathbf{\Delta}_\mathbf{V}\end{bmatrix} \in \mathbb{R}^{(n+m)\times r}$ the following holds:*

$$\left|\nabla^2 G(\mathbf{W})[\mathbf{\Delta}, \mathbf{\Delta}] - \nabla^2 g(\mathbf{W})[\mathbf{\Delta}, \mathbf{\Delta}]\right| \le 2c \left\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\right\|_F \left\|\mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F + c\left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F^2.$$

Proof of Lemma E.12.3. First note that

$$\begin{aligned}
&\nabla^2 G(\mathbf{W})[\mathbf{\Delta}, \mathbf{\Delta}] - \nabla^2 g(\mathbf{W})[\mathbf{\Delta}, \mathbf{\Delta}] \\
&= 2\left\langle \nabla f(\mathbf{X}), \mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\rangle - 2\left\langle \mathbf{X} - \mathbf{X}^\star, \mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\rangle + [\nabla^2 f(\mathbf{X})](\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top) - \left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F^2.
\end{aligned}$$

Now utilizing Lemma E.12.1 and (6.6), we have

$$\begin{aligned}\left|\nabla^2 G(\mathbf{W})[\mathbf{\Delta},\mathbf{\Delta}] - \nabla^2 g(\mathbf{W})[\mathbf{\Delta},\mathbf{\Delta}]\right| &\le 2\left|\left\langle \nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^\star), \mathbf{\Delta_U}\mathbf{\Delta_V}^\top\right\rangle - \langle \mathbf{X}-\mathbf{X}^\star, \mathbf{\Delta_U}\mathbf{\Delta_V}^\top\rangle\right| \\ &\quad + \left|[\nabla^2 f(\mathbf{X})](\mathbf{\Delta_U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta_V}^\top) - \left\|\mathbf{\Delta_U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta_V}^\top\right\|_F^2\right| \\ &\le 2c\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\right\|_F \left\|\mathbf{\Delta_U}\mathbf{\Delta_V}^\top\right\|_F + c\left\|\mathbf{\Delta_U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta_V}^\top\right\|_F^2.\end{aligned}$$

□

We provide one more result before proceeding to prove the main theorem.

Lemma E.12.4. *[125, Lemma E.1] Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two* $n \times r$ *matrices such that* $\mathbf{A}^\top\mathbf{B} = \mathbf{B}^\top\mathbf{A}$ *is PSD. Then*

$$\left\|(\mathbf{A}-\mathbf{B})\,\mathbf{A}^\top\right\|_F^2 \le \frac{1}{2(\sqrt{2}-1)}\left\|\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\right\|_F^2.$$

E.12.1 Local descent condition for the region $\mathcal{R}_1$

Similar to what used in Appendix E.11.1, we perform the change of variable $\mathbf{W}^\star\mathbf{R} \to \mathbf{W}^\star$ to avoid $\mathbf{R}$ in the following equations. With this change of variable we have instead $\mathbf{W}^\top\mathbf{W}^\star = \mathbf{W}^{\star\top}\mathbf{W}$ is PSD.

We first control $|\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W}-\mathbf{W}^\star\rangle|$ as follows:

$$\begin{aligned}|\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W}-\mathbf{W}^\star\rangle| &\le \left|\langle \nabla f(\mathbf{X}), (\mathbf{U}-\mathbf{U}^\star)\mathbf{V}^\top\rangle - \langle \mathbf{X}-\mathbf{X}^\star, (\mathbf{U}-\mathbf{U}^\star)\mathbf{V}^\top\rangle\right| \\ &\quad + \left|\langle \nabla f(\mathbf{X}), \mathbf{U}(\mathbf{V}-\mathbf{V}^\star)^\top\rangle - \langle \mathbf{X}-\mathbf{X}^\star, \mathbf{U}(\mathbf{V}-\mathbf{V}^\star)^\top\rangle\right| \\ &\le c\left\|\mathbf{X}-\mathbf{X}^\star\right\|_F\left(\|(\mathbf{U}-\mathbf{U}^\star)\mathbf{V}^\top\|_F + \|\mathbf{U}(\mathbf{V}-\mathbf{V}^\star)^\top\|_F\right) \\ &\le c\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\|_F\left\|\mathbf{W}(\mathbf{W}-\mathbf{W}^\star)^\top\right\|_F \\ &\le \frac{c}{2(\sqrt{2}-1)}\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\|_F^2\end{aligned}$$

where the second inequality utilizes $\nabla f(\mathbf{X}^\star) = \mathbf{0}$ and Lemma E.12.1, and the last inequality follows from Lemma E.12.4. The above result along with (E.31)- (E.32) gives

$$\begin{aligned}\langle \nabla G(\mathbf{W}), \mathbf{W}-\mathbf{W}^\star\rangle &\ge \langle \nabla g(\mathbf{W}), \mathbf{W}-\mathbf{W}^\star\rangle - |\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W}-\mathbf{W}^\star\rangle| \\ &\ge \langle \nabla g(\mathbf{W}), \mathbf{W}-\mathbf{W}^\star\rangle - \frac{c}{2(\sqrt{2}-1)}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2 \\ &\ge \frac{1}{16}\sigma_r(\mathbf{X}^\star)\,\mathrm{dist}^2(\mathbf{W},\mathbf{W}^\star) + \frac{1}{32}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2 \\ &\quad + \frac{1}{4\|\mathbf{X}^\star\|}\left\|\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2 \\ &\quad - \frac{c}{2(\sqrt{2}-1)}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2 \\ &\ge \frac{1}{16}\sigma_r(\mathbf{X}^\star)\,\mathrm{dist}^2(\mathbf{W},\mathbf{W}^\star) + \frac{1}{160}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2 \\ &\quad + \frac{1}{4\|\mathbf{X}^\star\|}\left\|\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2\end{aligned} \tag{E.52}$$

where we utilize $c \leq \frac{1}{50}$.

On the other hand, we control $\|\nabla G(\mathbf{W})\|_F$ with Lemma E.12.2 controlling the deviation between $\nabla G(\mathbf{W})$ and $\nabla g(\mathbf{W})$ as follows:

$$\begin{aligned}
&\|\nabla G(\mathbf{W})\|_F^2 = \|\nabla g(\mathbf{W}) + \nabla G(\mathbf{W}) - \nabla g(\mathbf{W})\|_F^2 \\
&\leq \frac{20}{19}\|\nabla g(\mathbf{W})\|_F^2 + 20\|\nabla g(\mathbf{W}) - \nabla G(\mathbf{W})\|_F^2 \\
&\leq \frac{20}{19}\|\nabla g(\mathbf{W})\|_F^2 + 20c^2\|\mathbf{W}\|^2\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2 \\
&= \frac{5}{19}\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W} + \widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2 + 20c^2\|\mathbf{W}\|^2\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F^2 \\
&\leq \left(\frac{5}{19}\frac{100}{99} + 20c^2\right)\left\|\left(\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right)\mathbf{W}\right\|_F^2 + 25\left\|\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\right\|_F^2 \\
&\leq (\frac{5}{19}\frac{100}{99} + 50c^2)(\sqrt{2}+1)^2\|\mathbf{X}^\star\|\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\|_F^2 + 25\|\widehat{\mathbf{W}}^\star\widehat{\mathbf{W}}^{\star\top}\mathbf{W}\|_F^2,
\end{aligned} \tag{E.53}$$

where the first inequality holds since $(a+b)^2 \leq \frac{1+\epsilon}{\epsilon}a^2 + (1+\epsilon)b^2$ for any $\epsilon > 0$, and the fourth line follows from (E.30).

Now combining (E.52)- (E.53) and assuming $c \leq \frac{1}{50}$ gives

$$\langle \nabla G(\mathbf{W}), \mathbf{W} - \mathbf{W}^\star\rangle \geq \frac{1}{16}\sigma_r(\mathbf{X}^\star)\operatorname{dist}^2(\mathbf{W}, \mathbf{W}^\star) + \frac{1}{260\|\mathbf{X}^\star\|}\|\nabla G(\mathbf{W})\|_F^2.$$

This completes the proof of (6.15).

E.12.2 Negative curvature for the region $\mathcal{R}_2$

Let $\mathbf{\Delta} = \mathbf{q}_k\boldsymbol{\alpha}^\top$ be defined as in (E.38). First note that

$$\begin{aligned}
\left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F^2 &\leq 2\left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top\right\|_F^2 + 2\left\|\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F^2 \\
&\leq 2\left\|\mathbf{W}\mathbf{\Delta}^\top\right\|_F^2 = 2\sigma_r^2(\mathbf{W}) \leq \sigma_r(\mathbf{X}^\star),
\end{aligned}$$

where the last equality holds because $\sigma_r(\mathbf{W}) \leq \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\mathbf{X}^\star)$. Also utilizing the particular structure in $\mathbf{\Delta}$ yields

$$\left\|\mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F = \frac{1}{2}\left\|\boldsymbol{\phi}_k\boldsymbol{\psi}_k^\top\right\|_F = \frac{1}{2}.$$

Due to the assumption $\frac{20}{19}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F \geq \|\mathbf{W}\mathbf{W}^\top\|_F$, we have

$$\begin{aligned}
\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\|_F &\leq \frac{\sqrt{2}}{2}\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\|_F \\
&\leq \frac{\sqrt{2}}{2}(\frac{20}{19}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F + \|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F) = \frac{39\sqrt{2}}{19}\|\mathbf{X}^\star\|_F.
\end{aligned}$$

Now combining the above results with Lemma E.12.3, we have

$$
\begin{aligned}
\nabla^2 G(\mathbf{W})[\mathbf{\Delta},\mathbf{\Delta}] &\le \nabla^2 g(\mathbf{W})[\mathbf{\Delta},\mathbf{\Delta}] + \left|\nabla^2 G(\mathbf{W})[\mathbf{\Delta},\mathbf{\Delta}] - \nabla^2 g(\mathbf{W})[\mathbf{\Delta},\mathbf{\Delta}]\right| \\
&\le -\frac{1}{4}\sigma_r(\mathbf{X}^\star) + 2c\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\right\|_F \left\|\mathbf{\Delta}_\mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F \\
&\quad + c\left\|\mathbf{\Delta}_\mathbf{U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta}_\mathbf{V}^\top\right\|_F^2 \\
&\le -\frac{1}{4}\sigma_r(\mathbf{X}^\star) + \frac{39}{19}\sqrt{2}c\|\mathbf{X}^\star\|_F + c\sigma_r(\mathbf{X}^\star) \\
&\le -\frac{1}{6}\sigma_r(\mathbf{X}^\star),
\end{aligned}
$$

where the last line holds when $c \le \frac{\sigma_r(\mathbf{X}^\star)}{50\|\mathbf{X}^\star\|_F}$. This completes the proof of (6.16).

E.12.3 Large gradient for the region $\mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$

To show that $G(\mathbf{W})$ has large gradient in these three regions, we mainly utilize Lemma E.12.2 to guarantee that $\nabla G(\mathbf{W})$ is close to $\nabla g(\mathbf{W})$.

E.12.3.1 Large gradient for the region $\mathcal{R}_3'$

Utilizing Lemma E.12.2, we have

$$
\begin{aligned}
&\|\nabla G(\mathbf{W})\|_F \\
&\ge \|\nabla g(\mathbf{W})\|_F - \|\nabla G(\mathbf{W}) - \nabla g(\mathbf{W})\|_F \\
&\ge \|\nabla g(\mathbf{W})\|_F - c\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F \|\mathbf{W}\| \\
&\ge \|\nabla g(\mathbf{W})\|_F - c(\frac{10}{9}\|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F + \|\mathbf{W}^\star\mathbf{W}^{\star\top}\|_F)\|\mathbf{W}\| \\
&\ge \frac{1}{10}\sigma_r^{3/2}(\mathbf{X}^\star) - c\frac{19}{9}2\|\mathbf{X}^\star\|_F\frac{20}{19}\sqrt{2}\|\mathbf{X}^\star\|^{1/2} \\
&\ge \frac{1}{27}\sigma_r^{3/2}(\mathbf{X}^\star),
\end{aligned}
$$

where the fourth line follows because $\left\|\mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F = 2\|\mathbf{X}^\star\|_F$ and $\|\mathbf{W}\| \le \frac{20}{19}\sqrt{2}\|\mathbf{X}^\star\|^{1/2}$, and the last line holds if $c \le \frac{1}{100}\frac{\sigma_r^{3/2}(\mathbf{X}^\star)}{\|\mathbf{X}^\star\|_F\|\mathbf{X}^\star\|^{1/2}}$. This completes the proof of (6.17).

E.12.3.2 Large gradient for the region $\mathcal{R}_3''$

Utilizing Lemma E.12.2 again, we have

$$
\begin{aligned}
&\|\nabla G(\mathbf{W})\|_F \\
&\geq \|\nabla g(\mathbf{W})\|_F - c\left(\left\|\mathbf{W}\mathbf{W}^\top\right\|_F + \left\|\mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F\right)\|\mathbf{W}\| \\
&\geq \frac{39}{800}\|\mathbf{W}\|^3 - c\left(\frac{10}{9}\left\|\mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F + \left\|\mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F\right)\|\mathbf{W}\| \\
&\geq \frac{39}{800}\|\mathbf{W}\|^3 - c\frac{19}{9}2\left\|\mathbf{X}^\star\right\|_F\|\mathbf{W}\| \\
&\geq \frac{39}{800}\|\mathbf{W}\|^3 - \frac{19}{450}\left\|\mathbf{X}^\star\right\|\|\mathbf{W}\| \\
&\geq \frac{1}{50}\|\mathbf{W}\|^3,
\end{aligned}
$$

where the fourth line holds if $c \leq \frac{1}{100}\frac{\sigma_r^{3/2}(\mathbf{X}^\star)}{\|\mathbf{X}^\star\|_F\|\mathbf{X}^\star\|^{1/2}}$ and the last follows from the fact that

$$
\|\mathbf{W}\| > \frac{20}{19}\|\mathbf{W}^\star\| \geq \frac{20}{19}\sqrt{2}\|\mathbf{X}^\star\|^{1/2}.
$$

This completes the proof of (6.18).

E.12.3.3 Large gradient for the region $\mathcal{R}_3'''$

To show (6.19), we first control $|\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W}\rangle|$ as follows:

$$
\begin{aligned}
&|\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W}\rangle| \\
&= 2\left|\left\langle \nabla f(\mathbf{U}\mathbf{V}^\top), \mathbf{U}\mathbf{V}^\top\right\rangle - \left\langle \mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star, \mathbf{U}\mathbf{V}^\top\right\rangle\right| \\
&\leq 2c\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\right\|_F\left\|\mathbf{U}\mathbf{V}^\top\right\|_F \\
&\leq 2c\frac{19}{20}\sqrt{2}\|\mathbf{W}\mathbf{W}^\top\|_F\frac{1}{2}\|\mathbf{W}\mathbf{W}^\top\|_F = \frac{19}{20}\sqrt{2}c\|\mathbf{W}\mathbf{W}^\top\|_F^2,
\end{aligned}
$$

where the first inequality utilizes the fact $\nabla f(\mathbf{X}^\star) = \mathbf{0}$ and Lemma E.12.1, and the last inequality holds because

$$
\begin{aligned}
\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\right\|_F &\leq \frac{\sqrt{2}}{2}\left\|\mathbf{W}\mathbf{W}^\top - \mathbf{W}^\star\mathbf{W}^{\star\top}\right\|_F \\
&\leq \frac{\sqrt{2}}{2}\left(\frac{9}{10}\left\|\mathbf{W}\mathbf{W}^\top\right\|_F + \left\|\mathbf{W}\mathbf{W}^\top\right\|_F\right) \\
&= \frac{19\sqrt{2}}{20}\left\|\mathbf{W}\mathbf{W}^\top\right\|_F
\end{aligned}
$$

and

$$
\begin{aligned}
\|\mathbf{W}\mathbf{W}^\top\|_F^2 &= \|\mathbf{U}\mathbf{U}^\top\|_F^2 + \|\mathbf{V}\mathbf{V}^\top\|_F^2 + 2\|\mathbf{U}\mathbf{V}^\top\|_F^2 \\
&\geq 4\|\mathbf{U}\mathbf{V}^\top\|_F^2
\end{aligned}
$$

by noting that

$$
\|\mathbf{U}\mathbf{U}^\top\|_F^2 + \|\mathbf{V}\mathbf{V}^\top\|_F^2 - 2\|\mathbf{U}\mathbf{V}^\top\|_F^2 = \|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 \geq 0.
$$

Now utilizing (E.48) to provide a lower bound for $\langle \nabla g(\mathbf{W}), \mathbf{W} \rangle$, we have

$$\begin{aligned}
&|\langle \nabla G(\mathbf{W}), \mathbf{W} \rangle| \\
&\geq \langle \nabla g(\mathbf{W}), \mathbf{W} \rangle - |\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W} \rangle| \\
&> \frac{1}{20} \|\mathbf{W}\mathbf{W}^\top\|_F^2 - \frac{19}{20}\sqrt{2}c \|\mathbf{W}\mathbf{W}^\top\|_F^2 \\
&\geq \frac{1}{45} \|\mathbf{W}\mathbf{W}^\top\|_F^2,
\end{aligned}$$

where the last line holds when $c \leq \frac{1}{50}$. Thus,

$$\|\nabla G(\mathbf{W})\|_F \geq \frac{1}{\|\mathbf{W}\|} |\langle \nabla G(\mathbf{W}), \mathbf{W} \rangle| > \frac{1}{45} \|\mathbf{W}\mathbf{W}^\top\|_F^{3/2},$$

where we utilize $\|\mathbf{W}\| \leq \left(\|\mathbf{W}\mathbf{W}^\top\|_F\right)^{1/2}$. This completes the proof of (6.19).

□

APPENDIX F

APPENDICES FOR CHAPTER 8

F.1 Proof of Proposition 8.2.1

Proposition F.1.1 (Proposition 8.2.1)**.** *Let $f(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{J} f_j(\mathbf{x},\mathbf{y}_j)$ be an objective function as in* (8.3) *and let $g(\mathbf{z})$ be as in* (8.7) *with $\mathbf{z} = (\mathbf{x}^1,\dots,\mathbf{x}^J,\mathbf{y}_1,\dots,\mathbf{y}_J)$. Suppose that each f_j has Lipschitz gradient, i.e., ∇f_j is Lipschitz continuous with constant $L_j > 0$. Then ∇g is Lipschitz continuous with constant*

$$L_g = L + \frac{2\omega}{\mu},$$

where $L := \max_j L_j$, $\omega := \sum_{i\neq j}^{J} \widetilde{w}_{ji}$, and $\widetilde{w}_{ji}$ and μ are the DGD+LOCAL weights and stepsize as in (8.5)*.*

Proof. Let $L = \max_j L_j$ and

$$\boldsymbol{\delta}_{\mathbf{z}} = (\boldsymbol{\delta}_{\mathbf{x}^1},\dots,\boldsymbol{\delta}_{\mathbf{x}^J},\boldsymbol{\delta}_{\mathbf{y}_1},\dots,\boldsymbol{\delta}_{\mathbf{y}_J}).$$

First, for any $\mathbf{z}$ and $\boldsymbol{\delta}_{\mathbf{z}}$, and using the symmetry of $\mathbf{W} = \{w_{ij}\}$, we have

$$\nabla g(\mathbf{z}+\boldsymbol{\delta}_{\mathbf{z}}) - \nabla g(\mathbf{z}) = \begin{bmatrix} \nabla_{\mathbf{x}} f_1(\mathbf{x}^1+\boldsymbol{\delta}_{\mathbf{x}^1},\mathbf{y}_1+\boldsymbol{\delta}_{\mathbf{y}_1}) - \nabla_{\mathbf{x}} f_1(\mathbf{x}^1,\mathbf{y}_1) + 4\sum_{i=1}^{J} w_{1i}(\boldsymbol{\delta}_{\mathbf{x}^1} - \boldsymbol{\delta}_{\mathbf{x}^i}) \\ \vdots \\ \nabla_{\mathbf{x}} f_J(\mathbf{x}^J+\boldsymbol{\delta}_{\mathbf{x}^J},\mathbf{y}_J+\boldsymbol{\delta}_{\mathbf{y}_J}) - \nabla_{\mathbf{x}} f_J(\mathbf{x}^J,\mathbf{y}_J) + 4\sum_{i=1}^{J} w_{Ji}(\boldsymbol{\delta}_{\mathbf{x}^J} - \boldsymbol{\delta}_{\mathbf{x}^i}) \\ \nabla_{\mathbf{y}} f_1(\mathbf{x}^1+\boldsymbol{\delta}_{\mathbf{x}^1},\mathbf{y}_1+\boldsymbol{\delta}_{\mathbf{y}_1}) - \nabla_{\mathbf{y}} f_1(\mathbf{x}^1,\mathbf{y}_1) \\ \vdots \\ \nabla_{\mathbf{y}} f_J(\mathbf{x}^J+\boldsymbol{\delta}_{\mathbf{x}^J},\mathbf{y}_J+\boldsymbol{\delta}_{\mathbf{y}_J}) - \nabla_{\mathbf{y}} f_J(\mathbf{x}^J,\mathbf{y}_J) \end{bmatrix}$$

Then with some rearrangement, denoting $\nabla f_j = \nabla_{\left[\begin{smallmatrix}\mathbf{x}\\ \mathbf{y}\end{smallmatrix}\right]} f_j$ and using the triangle inequality, we can obtain

$$\begin{aligned} \|\nabla g(\mathbf{z}+\boldsymbol{\delta}_{\mathbf{z}}) - \nabla g(\mathbf{z})\|_2 &\leq \left\| \begin{bmatrix} \nabla f_1(\mathbf{x}^1+\boldsymbol{\delta}_{\mathbf{x}^1},\mathbf{y}_1+\boldsymbol{\delta}_{\mathbf{y}_1}) - \nabla f_1(\mathbf{x}^1,\mathbf{y}_1) \\ \vdots \\ \nabla f_J(\mathbf{x}^J+\boldsymbol{\delta}_{\mathbf{x}^J},\mathbf{y}_J+\boldsymbol{\delta}_{\mathbf{y}_J}) - \nabla f_J(\mathbf{x}^J,\mathbf{y}_J) \end{bmatrix} \right\|_2 + 4 \left\| \begin{bmatrix} \sum_{i=1}^{J} w_{1i}(\boldsymbol{\delta}_{\mathbf{x}^1} - \boldsymbol{\delta}_{\mathbf{x}^i}) \\ \vdots \\ \sum_{i=1}^{J} w_{Ji}(\boldsymbol{\delta}_{\mathbf{x}^J} - \boldsymbol{\delta}_{\mathbf{x}^i}) \end{bmatrix} \right\|_2 \\ &\leq \sqrt{\sum_{j=1}^{J} L_j^2 \left\| \left[\begin{smallmatrix} \boldsymbol{\delta}_{\mathbf{x}^j} \\ \boldsymbol{\delta}_{\mathbf{y}_j} \end{smallmatrix}\right] \right\|_2^2} + 4\sqrt{\sum_{j=1}^{J} \left(\sum_{i=1}^{J} w_{ji} \right)^2 \|\boldsymbol{\delta}_{\mathbf{x}^j}\|_2^2} + 4\sqrt{\sum_{j=1}^{J} \left\| \sum_{i=1}^{J} w_{ji}\boldsymbol{\delta}_{\mathbf{x}^i} \right\|_2^2} \\ &\leq L\|\boldsymbol{\delta}_{\mathbf{z}}\|_F + \left(4\max_j \sum_{i=1}^{J} w_{ji} \right) \left\| \begin{bmatrix} \boldsymbol{\delta}_{\mathbf{x}^1} & \cdots & \boldsymbol{\delta}_{\mathbf{x}^1} \end{bmatrix} \right\|_F + 4 \left(\max_j \sum_{i=1}^{J} w_{ji} \right) \left\| \begin{bmatrix} \boldsymbol{\delta}_{\mathbf{x}^1} & \cdots & \boldsymbol{\delta}_{\mathbf{x}^1} \end{bmatrix} \right\|_F. \end{aligned}$$

where in the last line we use

$$\sqrt{\sum_{j=1}^{J}\left\|\sum_{i=j}^{J} w_{ji}\boldsymbol{\delta}_{\mathbf{x}^i}\right\|_2^2} = \left\|\begin{bmatrix}\boldsymbol{\delta}_{\mathbf{x}^1} & \cdots & \boldsymbol{\delta}_{\mathbf{x}^1}\end{bmatrix}\mathbf{W}\right\|_F = \left\|\mathbf{W}^\top\begin{bmatrix}\boldsymbol{\delta}_{\mathbf{x}^1} & \cdots & \boldsymbol{\delta}_{\mathbf{x}^1}\end{bmatrix}^\top\right\|_F$$
$$\leq \|\mathbf{W}\|\left\|\begin{bmatrix}\boldsymbol{\delta}_{\mathbf{x}^1} & \cdots & \boldsymbol{\delta}_{\mathbf{x}^1}\end{bmatrix}\right\|_F$$
$$\leq \left(\max_j \sum_{i=1}^{J} w_{ji}\right)\left\|\begin{bmatrix}\boldsymbol{\delta}_{\mathbf{x}^1} & \cdots & \boldsymbol{\delta}_{\mathbf{x}^1}\end{bmatrix}\right\|_F$$

since $\|\mathbf{W}\| \leq \max_j \sum_{i\neq j} w_{ji} = \max_j \sum_{i=1}^{J} w_{ji}$ in view of that $\mathbf{W}$ is symmetric, $w_{ii} = 0$ and $w_{ij} \geq 0$ by (8.6).

Finally, using the definition of w_{ji} (8.6), we have $\max_j \sum_{i=1}^{J} w_{ji} = \max_j \sum_{i\neq j}^{J} w_{ji} = \max_j \frac{\sum_{i\neq j}^{J} \widetilde{w}_{ji}}{4\mu} =: \frac{\omega}{4\mu}$, and further by the inequality $\left\|\begin{bmatrix}\boldsymbol{\delta}_{\mathbf{x}^1} & \cdots & \boldsymbol{\delta}_{\mathbf{x}^1}\end{bmatrix}\right\|_F \leq \|\boldsymbol{\delta}_Z\|_F$, we obtain that ∇g is Lipschitz continuous with constant

$$L_g = L + 4\left(\frac{\omega}{4\mu}\right) + 4\left(\frac{\omega}{4\mu}\right) = L + \frac{2\omega}{\mu}.$$

□

F.2 Proof of Theorem 8.2.4

Theorem F.2.1 (Theorem 8.2.4)**.** *Let $\rho > 0$, and consider an objective function h where:*

1. $\inf_{\mathbb{R}^n} h > -\infty$,
2. *h satisfies the Łojasiewicz inequality within B_ρ,*
3. *h is twice-continuously differentiable, and*
4. *$|h(\mathbf{x})| \leq L_0$, $\|\nabla h(\mathbf{x})\| \leq L_1$, and $\left\|\nabla^2 h(\mathbf{x})\right\|_2 \leq L_2$ for all $\mathbf{x} \in B_{2\rho}$.*

Suppose the gradient descent stepsize

$$\mu < \frac{1}{L_2 + \frac{4L_1}{\rho} + \frac{(4+2\pi)L_0}{\rho^2}}. \tag{8.8}$$

Suppose $\mathbf{x}(0)$ is chosen randomly from a probability distribution supported on a set $S \subseteq B_\rho$ with S having positive measure, and suppose that under such random initialization, there is a positive probability that the sequence $\{\mathbf{x}(k)\}$ remains bounded in B_ρ and all limit points of $\{\mathbf{x}(k)\}$ are in B_ρ.

Then conditioned on observing that $\{\mathbf{x}(k)\} \subseteq B_\rho$ and all limit points of $\{\mathbf{x}(k)\}$ are in B_ρ, gradient descent converges to a critical point of h, and the probability that this critical point is a strict saddle point is zero.

Proof. The proof involves constructing a function $\widetilde{h}$ such that $\widetilde{h}(\mathbf{x}) = h(\mathbf{x})$ for all $\mathbf{x} \in B_\rho$ but where $\widetilde{h}$ has a globally Lipschitz gradient.

To do this, first define a window function $w : \mathbb{R}^n \to \mathbb{R}$,

$$w(\mathbf{x}) = \begin{cases} 1, & \|\mathbf{x}\| \le \rho \\ 2 - \frac{\|\mathbf{x}\|}{\rho} + \frac{1}{2\pi}\sin\left(\frac{2\pi\|\mathbf{x}\|}{\rho}\right), & \rho < \|\mathbf{x}\| < 2\rho \\ 0, & \|\mathbf{x}\| \ge 2\rho, \end{cases}$$

where $\|\cdot\| = \|\cdot\|_2$. Note also that

$$\nabla w(\mathbf{x}) = \begin{cases} 0, & \|\mathbf{x}\| \le \rho \\ -\frac{2\mathbf{x}}{\rho\|\mathbf{x}\|}\sin^2\left(\frac{\pi\|\mathbf{x}\|}{\rho}\right), & \rho < \|\mathbf{x}\| < 2\rho \\ 0, & \|\mathbf{x}\| \ge 2\rho \end{cases}$$

and

$$\nabla^2 w(\mathbf{x}) = \begin{cases} 0, & \|\mathbf{x}\| \le \rho \\ -\frac{2}{\rho\|\mathbf{x}\|}\sin^2\left(\frac{\pi\|\mathbf{x}\|}{\rho}\right) I + \left(\frac{2}{\rho\|\mathbf{x}\|^3}\sin^2\left(\frac{\pi\|\mathbf{x}\|}{\rho}\right) - \frac{2\pi}{\rho^2\|\mathbf{x}\|^2}\sin\left(\frac{2\pi\|\mathbf{x}\|}{\rho}\right)\right)\mathbf{x}\mathbf{x}^\top, & \rho < \|\mathbf{x}\| < 2\rho \\ 0, & \|\mathbf{x}\| \ge 2\rho \end{cases},$$

where I denotes the n-by-n identity matrix. It is easy to verify that $w \in \mathcal{C}^2$ and $|w(\mathbf{x})| \le 1$. To bound the gradient ∇w, we have

$$\|\nabla w\| = \left\| -\frac{2\mathbf{x}}{\rho\|\mathbf{x}\|}\sin^2\left(\frac{\pi\|\mathbf{x}\|}{\rho}\right)\right\| \le \frac{2}{\rho}.$$

For the Hessian $\nabla^2 w$, we have

$$\left\|\nabla^2 w\right\| \le \left\|\frac{2}{\rho\|\mathbf{x}\|}\sin^2\left(\frac{\pi\|\mathbf{x}\|}{\rho}\right) I\right\| + \left\|\left(\frac{2}{\rho\|\mathbf{x}\|^3}\sin^2\left(\frac{\pi\|\mathbf{x}\|}{\rho}\right) - \frac{2\pi}{\rho^2\|\mathbf{x}\|^2}\sin\left(\frac{2\pi\|\mathbf{x}\|}{\rho}\right)\right)\mathbf{x}^\top\mathbf{x}\right\| \le \frac{4+2\pi}{\rho^2}.$$

Now, define

$$\widetilde{h}(\mathbf{x}) = h(\mathbf{x})\, w(\mathbf{x}) = \begin{cases} h(\mathbf{x}), & \|\mathbf{x}\| \le \rho \\ h(\mathbf{x})\left(2 - \frac{\|\mathbf{x}\|}{\rho} + \frac{1}{2\pi}\sin\left(\frac{2\pi\|\mathbf{x}\|}{\rho}\right)\right), & \rho < \|\mathbf{x}\| < 2\rho \\ 0, & \|\mathbf{x}\| \ge 2\rho. \end{cases}$$

We have the following properties for $\widetilde{h}$:

- Since $h = \widetilde{h}$ in B_ρ, $\widetilde{h}$ satisfies the Łojasiewicz inequality in B_ρ.
- Since $h, w \in C^2$, $\widetilde{h} \in C^2$.
- Since $\inf_{\mathbb{R}^n} h > -\infty$ and $\inf_{\mathbb{R}^n} w > -\infty$, $\inf_{\mathbb{R}^n} \widetilde{h} > -\infty$.
- To globally bound the Lipschitz constant of the gradient of $\widetilde{h}$, note that

$$\begin{aligned} \left\|\nabla^2\widetilde{h}\right\| &= \left\|w\cdot\nabla^2 h + \nabla h\cdot(\nabla w)^\top + \nabla w\cdot(\nabla h)^\top + h\cdot\nabla^2 w\right\| \\ &\le |w|\left\|\nabla^2 h\right\| + 2\|\nabla w\|\|\nabla h\| + |h|\left\|\nabla^2 w\right\| \\ &\le L_2 + \frac{4L_1}{\rho} + \frac{(4+2\pi)L_0}{\rho^2}. \end{aligned}$$

Now consider the gradient descent algorithm with stepsize μ satisfying (8.8). Define

$$T_h = \{\mathbf{x}(0) \in B_\rho : \text{all } \{\mathbf{x}(k)\} \subseteq B_\rho \text{ and all limit points of } \{\mathbf{x}(k)\} \text{ are in } B_\rho \\ \text{when gradient descent is run on } h \text{ starting at } \mathbf{x}(0)\}$$

and

$$T_{\widetilde{h}} = \{\mathbf{x}(0) \in B_\rho : \text{all } \{\mathbf{x}(k)\} \subseteq B_\rho \text{ and all limit points of } \{\mathbf{x}(k)\} \text{ are in } B_\rho \\ \text{when gradient descent is run on } \widetilde{h} \text{ starting at } \mathbf{x}(0)\}.$$

Similarly, define

$$\Sigma_h = \{\mathbf{x}(0) \in B_\rho : \ \{\mathbf{x}(k)\} \text{ converges to a strict saddle when gradient descent is run on } h \text{ starting at } \mathbf{x}(0)\}$$

and

$$\Sigma_{\widetilde{h}} = \{\mathbf{x}(0) \in B_\rho : \ \{\mathbf{x}(k)\} \text{ converges to a strict saddle when gradient descent is run on } \widetilde{h} \text{ starting at } \mathbf{x}(0)\}.$$

Using the above properties, we see that Theorem 8.2.2 can be applied to $\widetilde{h}$, and so we conclude that $\Sigma_{\widetilde{h}}$ has measure zero.

Now, after running gradient descent on h from a random initialization as in the theorem statement, condition on observing that $\{\mathbf{x}(k)\} \subseteq B_\rho$ and all limit points of $\{\mathbf{x}(k)\}$ are in B_ρ, i.e., that $\mathbf{x}(0) \in T_h$. Because $\{\mathbf{x}(k)\} \subseteq B_\rho$ and all limit points of $\{\mathbf{x}(k)\}$ are in B_ρ, and because $\{\mathbf{x}(k)\}$ matches the sequence that would be obtained by running gradient descent on $\widetilde{h}$, we can apply Theorem 8.2.3 to conclude that $\{\mathbf{x}(k)\}$ converges to a critical point of $\widetilde{h}$, and since this critical point belongs to B_ρ and $\widetilde{h} = h$ inside B_ρ, we conclude that this is also a critical point of h.

Finally, using the definition of conditional probability, we have

$$\begin{aligned} P(\mathbf{x}(0) \in \Sigma_h | \mathbf{x}(0) \in T_h) &= \frac{P(\mathbf{x}(0) \in \Sigma_h \cap T_h)}{P(\mathbf{x}(0) \in T_h)} \\ &= \frac{P(\mathbf{x}(0) \in \Sigma_{\widetilde{h}} \cap T_{\widetilde{h}})}{P(\mathbf{x}(0) \in T_h)}, \end{aligned}$$

where the second equality follows from the fact that $\widetilde{h} = h$ inside B_ρ: if a sequence of iterations stays bounded inside B_ρ and converges to a strict saddle when gradient descent is run on h, the same will hold when gradient descent is run on $\widetilde{h}$, and vice versa. Since $\Sigma_{\widetilde{h}}$ has zero measure and because $\mathbf{x}(0)$ is chosen randomly from a probability distribution supported on a set $S \subseteq B_\rho$ with S having positive measure, $P(\mathbf{x}(0) \in \Sigma_{\widetilde{h}} \cap T_{\widetilde{h}}) = 0$. Also, by assumption, $P(\mathbf{x}(0) \in T_h) > 0$. Therefore, $P(\mathbf{x}(0) \in \Sigma_h | \mathbf{x}(0) \in T_h) = \frac{0}{\text{nonzero}} = 0$.

□

F.3 Proof of Proposition 8.2.2

Proposition F.3.1 (Proposition 8.2.2)**.** *Let $f(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{J} f_j(\mathbf{x}, \mathbf{y}_j)$ be as in* (8.3)*. Suppose the topology defined by $\mathbf{W}$ is connected. Also suppose there exist $\mathbf{x}^\star$ (which is independent of j) and $\mathbf{y}_j^\star, j \in [J]$ such that*

$$(\mathbf{x}^\star, \mathbf{y}_j^\star) \in \arg\min_{\mathbf{x}, \mathbf{y}_j} f_j(\mathbf{x}, \mathbf{y}_j),\ \forall\, j \in [J]. \tag{8.12}$$

Then $g(\mathbf{z})$ defined in (8.7) *satisfies*

$$\min_{\mathbf{z}} g(\mathbf{z}) = \min_{\mathbf{x}, \mathbf{y}} f(\mathbf{x}, \mathbf{y}),$$

and $g(\mathbf{z})$ achieves its global minimum only for $\mathbf{z}$ with $\mathbf{x}^1 = \cdots = \mathbf{x}^J$.

Proof. First note that

$$\min_{\mathbf{z}} g(\mathbf{z}) = \sum_{j=1}^{J} \left(f_j(\mathbf{x}^j, \mathbf{y}_j) + \sum_{i=1}^{J} w_{ji} \|\mathbf{x}^j - \mathbf{x}^i\|_2^2 \right) \geq \sum_{j=1}^{J} \min_{\mathbf{x}^j, \mathbf{y}_j} f_j(\mathbf{x}^j, \mathbf{y}_j) = \sum_{j=1}^{J} f_j(\mathbf{x}^\star, \mathbf{y}_j^\star) = \min_{\mathbf{x}, \mathbf{y}} f(\mathbf{x}, \mathbf{y}). \tag{F.1}$$

On the other hand, we have

$$\begin{aligned}
\min_{\mathbf{z}} g(\mathbf{z}) &= \min_{\mathbf{z}} \sum_{j=1}^{J} \left(f_j(\mathbf{x}^j, \mathbf{y}_j) + \sum_{i=1}^{J} w_{ji} \|\mathbf{x}^j - \mathbf{x}^i\|_2^2 \right) \\
&\leq \min_{\mathbf{z}: \mathbf{x}^1 = \cdots = \mathbf{x}^J} \sum_{j=1}^{J} \left(f_j(\mathbf{x}^j, \mathbf{y}_j) + \sum_{i=1}^{J} w_{ji} \|\mathbf{x}^j - \mathbf{x}^i\|_2^2 \right) \\
&= \min_{\mathbf{x}, \mathbf{y}} \sum_{j=1}^{J} f_j(\mathbf{x}, \mathbf{y}_j) = \min_{\mathbf{x}, \mathbf{y}} f(\mathbf{x}, \mathbf{y}).
\end{aligned}$$

Thus, we have

$$\min_{\mathbf{z}} g(\mathbf{z}) = \min_{\mathbf{x}, \mathbf{y}} f(\mathbf{x}, \mathbf{y}).$$

The proof is completed by noting that (F.1) achieves the equality only at $\mathbf{z}$ with $\mathbf{x}^1 = \cdots = \mathbf{x}^J$ since the topology defined by $\mathbf{W}$ is connected. □

F.4 Proof of Proposition 8.2.3

Proposition F.4.1 (Proposition 8.2.3)**.** *Let $f(\mathbf{x}, \mathbf{y})$ be as in* (8.3) *and $g(\mathbf{z})$ be as in* (8.7) *with $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^J, \mathbf{y}_1, \ldots, \mathbf{y}_J)$. Suppose the matrix $\mathbf{W}$ is connected and symmetric. Also suppose the gradient of f_j satisfies the following symmetric property:*

$$\langle \nabla_{\mathbf{x}} f_j(\mathbf{x}, \mathbf{y}_j), \mathbf{x} \rangle = \langle \nabla_{\mathbf{y}_j} f_j(\mathbf{x}, \mathbf{y}_j), \mathbf{y}_j \rangle \tag{8.13}$$

for all $j \in [J]$. Then, any critical point of g must satisfy $\mathbf{x}^1 = \cdots = \mathbf{x}^J$.

Proof. The critical points of the objective function in (8.7) satisfy

$$\nabla_{\mathbf{x}^j} g(\mathbf{z}) = \nabla_{\mathbf{x}} f_j(\mathbf{x}^j, \mathbf{y}_j) + \sum_{i=1}^{J} 2w_{ji}(\mathbf{x}^j - \mathbf{x}^i) = \mathbf{0}, \tag{F.2}$$

$$\nabla_{\mathbf{y}^j} g(\mathbf{z}) = \nabla_{\mathbf{y}_j} f_j(\mathbf{x}^j, \mathbf{y}_j) = \mathbf{0}, \forall\, j \in [J]. \tag{F.3}$$

Now taking the inner product of both sides in (F.2) with $\mathbf{x}^j$ and also the inner product of both sides in (F.3) with $\mathbf{y}^j$ and using the property (8.13), we have

$$\sum_{i=1}^{J} 2w_{ji}\langle \mathbf{x}^j, \mathbf{x}^j - \mathbf{x}^i \rangle = 0$$

for all $j \in [J]$. Using the symmetric property of $\mathbf{W}$, we then have

$$\sum_{j=1}^{J}\sum_{i=1}^{J} w_{ji}\|\mathbf{x}^j - \mathbf{x}^i\|^2 = 0.$$

Therefore,

$$\mathbf{x}^i = \mathbf{x}^j, \text{ if } w_{ij} \neq 0$$

for any $i, j \in [J]$. Since the topology defined by $\mathbf{W}$ is connected, we finally have

$$\mathbf{x}^1 = \cdots = \mathbf{x}^J.$$

□

F.5 Proof of Theorem 8.2.7

Theorem F.5.1 (Theorem 8.2.7)**.** *Let $\mathcal{C}_f$ denote the set of critical points of* (8.3)*:*

$$\mathcal{C}_f := \{\mathbf{x}, \mathbf{y} : \nabla f(\mathbf{x}, \mathbf{y}) = \mathbf{0}\},$$

and let $\mathcal{C}_g$ denote the set of critical points of (8.7)*:*

$$\mathcal{C}_g := \left\{\mathbf{z} : \nabla g(\mathbf{z}) = \mathbf{0}\right\}.$$

Then, for any $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^J, \mathbf{y}) \in \mathcal{C}_g$ with $\mathbf{x}^1 = \cdots = \mathbf{x}^J = \mathbf{x}$, we have $(\mathbf{x}, \mathbf{y}) \in \mathcal{C}_f$. Furthermore, if $(\mathbf{x}, \mathbf{y})$ is a strict saddle of f, then $\mathbf{z} = (\mathbf{x}, \ldots, \mathbf{x}, \mathbf{y})$ is also a strict saddle of g.

Proof. We rewrite $\mathcal{C}_f$ as:

$$\mathcal{C}_f = \left\{ \mathbf{x}, \mathbf{y} : \sum_{j=1}^{J} \nabla_{\mathbf{x}} f_j(\mathbf{x}, \mathbf{y}_j) = \mathbf{0}, \nabla_{\mathbf{y}_j} f_j(\mathbf{x}, \mathbf{y}_j) = \mathbf{0}, \forall j \in [J] \right\}.$$

The critical points of the objective function in (8.7) satisfy

$$\begin{aligned}
\nabla_{\mathbf{x}^j} g(\mathbf{z}) &= \nabla_{\mathbf{x}} f_j(\mathbf{x}^j, \mathbf{y}_j) + \sum_{i=1}^{J} 2(w_{ij} + w_{ji})(\mathbf{x}^j - \mathbf{x}^i) = \mathbf{0}, \\
\nabla_{\mathbf{y}^j} g(\mathbf{z}) &= \nabla_{\mathbf{y}_j} f_j(\mathbf{x}^j, \mathbf{y}_j) = \mathbf{0}, \forall\, j \in [J].
\end{aligned}$$

With this, we rewrite $\mathcal{C}_g$ as

$$\mathcal{C}_g = \left\{ \mathbf{z} : \nabla_{\mathbf{x}} f_j(\mathbf{x}^j, \mathbf{y}_j) + \sum_{i=1}^{J} 2(w_{ij} + w_{ji})(\mathbf{x}^j - \mathbf{x}^i) = \mathbf{0}, \nabla_{\mathbf{y}_j} f_j(\mathbf{x}^j, \mathbf{y}_j) = \mathbf{0}, \forall\, j \in [J] \right\}.$$

Thus, for any $\mathbf{z} = (\mathbf{x}^1, \ldots, \mathbf{x}^J, \mathbf{y}) \in \mathcal{C}_g$ with $\mathbf{x}^1 = \cdots = \mathbf{x}^J = \mathbf{x}$, we have that $(\mathbf{x}, \mathbf{y})$ is a critical point of (8.3), i.e., $(\mathbf{x}, \mathbf{y}) \in \mathcal{C}_f$. In what follows, we check how the Hessian information (especially the smallest eigenvalue of the Hessian) of $(\mathbf{x}, \mathbf{y})$ is transformed to $\mathbf{z}$.

At any point $(\mathbf{x}, \mathbf{y})$, the Hessian quadratic form of f for any $\mathbf{q}_{\mathbf{x}}$ and $\mathbf{q}_{\mathbf{y}} = \begin{bmatrix} \mathbf{q}_{\mathbf{y}_1}^\top & \cdots & \mathbf{q}_{\mathbf{y}_J}^\top \end{bmatrix}^\top$ is given by

$$[\nabla^2 f(\mathbf{x}, \mathbf{y})](\begin{bmatrix} \mathbf{q}_{\mathbf{x}} \\ \mathbf{q}_{\mathbf{y}} \end{bmatrix}, \begin{bmatrix} \mathbf{q}_{\mathbf{x}} \\ \mathbf{q}_{\mathbf{y}} \end{bmatrix}) = \sum_{j=1}^{J} \nabla^2 f_j(\begin{bmatrix} \mathbf{q}_{\mathbf{x}} \\ \mathbf{q}_{\mathbf{y}_j} \end{bmatrix}, \begin{bmatrix} \mathbf{q}_{\mathbf{x}} \\ \mathbf{q}_{\mathbf{y}_j} \end{bmatrix}).$$

At any point $\mathbf{z}$, the Hessian quadratic form of g for any $\mathbf{q} = \begin{bmatrix} \mathbf{q}_{\mathbf{x}^1}^\top & \cdots & \mathbf{q}_{\mathbf{x}^J}^\top & \mathbf{q}_{\mathbf{y}_1}^\top & \cdots & \mathbf{q}_{\mathbf{y}_J}^\top \end{bmatrix}$ is given by

$$[\nabla^2 g(\mathbf{z})](\mathbf{q}, \mathbf{q}) = \sum_{j=1}^{J} \nabla^2 f_j(\begin{bmatrix} \mathbf{q}_{\mathbf{x}^j} \\ \mathbf{q}_{\mathbf{y}_j} \end{bmatrix}, \begin{bmatrix} \mathbf{q}_{\mathbf{x}^j} \\ \mathbf{q}_{\mathbf{y}_j} \end{bmatrix}) + \sum_{j=1}^{J} 2 w_{ji} \|\mathbf{q}_{\mathbf{x}^j} - \mathbf{q}_{\mathbf{x}^i}\|_2^2.$$

Now suppose $\lambda_{\min}(\nabla^2 f(\mathbf{x}, \mathbf{y})) < 0$ (where $\lambda_{\min}$ denotes the smallest eigenvalue), i.e., there exist $\mathbf{q}_{\mathbf{x}}, \mathbf{q}_{\mathbf{y}}$ such that $[\nabla^2 f(\mathbf{x}, \mathbf{y})](\begin{bmatrix} \mathbf{q}_{\mathbf{x}} \\ \mathbf{q}_{\mathbf{y}} \end{bmatrix}, \begin{bmatrix} \mathbf{q}_{\mathbf{x}} \\ \mathbf{q}_{\mathbf{y}} \end{bmatrix}) < 0$. Choosing $\mathbf{q}_{\mathbf{x}^1} = \cdots = \mathbf{q}_{\mathbf{x}^J} = \mathbf{q}_{\mathbf{x}}$, we have $[\nabla^2 g(\mathbf{z})](\mathbf{q}, \mathbf{q}) < 0$, i.e., $\lambda_{\min}(\nabla^2 g(\mathbf{z})) < 0$. □

F.6 Proof of Theorem 8.3.1

Theorem F.6.1 (Theorem 8.3.1)**.** *For any data matrix $\mathbf{Y}$, every critical point (i.e., every point where the gradient is zero) of problem* (8.14) *is either a global minimum or a strict saddle point, where the Hessian has at least one negative eigenvalue.*

Proof. Denote by $h(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\|\mathbf{U}\mathbf{V}^\top - \mathbf{Y}\|_F^2$. Let $\mathcal{C}$ denote the set of critical points of h:

$$\mathcal{C} = \left\{ (\mathbf{U}, \mathbf{V}) : (\mathbf{U}\mathbf{V}^\top - \mathbf{Y})\mathbf{V} = \mathbf{0},\ (\mathbf{U}\mathbf{V}^\top - \mathbf{Y})^\top \mathbf{U} = \mathbf{0} \right\}.$$

Our goal is to characterize the behavior of all the critical points that are not global minima. In particular, we want to show that every critical point of h is either a global minimum or a strict saddle. Towards that end, we first recall the following result concerning the degenerate critical points.

Lemma F.6.1. *[206, Theorem 8 with $\mathbf{X} = \mathbf{I}$] Any pair $(\mathbf{U}, \mathbf{V}) \in \mathcal{C}$ that is degenerate (i.e., $\operatorname{rank}(\mathbf{U}\mathbf{V}^\top) < r$) is either a global minimum of h (i.e., $\mathbf{U}\mathbf{V}^\top = \mathbf{Y}_r$ where $\mathbf{Y}_r$ is a rank-r approximation of $\mathbf{Y}$) or a strict saddle (i.e., $\lambda_{\min}(\nabla^2 h(\mathbf{U}, \mathbf{V})) < 0$).*

Note that the above result holds for any matrix $\mathbf{Y}$. When $\operatorname{rank}(\mathbf{Y}) \leq r$, then $\mathbf{Y}_r = \mathbf{Y}$. It follows from Lemma F.6.1 that the behavior of all degenerate critical points is quite clear. For the remaining non-degenerate critical points, using the same argument in cite[Theorems 2–4]nvx:zhu2017global, we first establish the following results concerning the critical points that are also balanced (i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}$).

Lemma F.6.2. *[93, Theorems 2–4] Any pair $(\mathbf{U}, \mathbf{V}) \in \mathcal{C}$ satisfying $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}$ is either a global minimum of h or a strict saddle.*

The above result also holds for any matrix $\mathbf{Y}$. With this result, we now show that non-degenerate critical points behave similarly to degenerate ones.

Lemma F.6.3. *Any pair $(\mathbf{U}, \mathbf{V}) \in \mathcal{C}$ that is non-degenerate (i.e., $\operatorname{rank}(\mathbf{U}\mathbf{V}^\top) = r$) is either a global minimum of h or a strict saddle.*

Proof of Lemma F.6.3. Suppose $(\mathbf{U}, \mathbf{V})$ is not a global minimum of h. Let $\mathbf{U}\mathbf{V}^\top = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top$ be a reduced SVD of $\mathbf{U}\mathbf{V}^\top$. Since $\operatorname{rank}(\mathbf{U}\mathbf{V}^\top) = r$ and both $\mathbf{U}$ and $\mathbf{V}$ have only r columns, we know $\operatorname{rank}(\mathbf{U}) = \operatorname{rank}(\mathbf{V}) = r$. Denote by $\mathbf{D} = (\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{P}\mathbf{\Sigma}^{1/2}$ and $\mathbf{G} = (\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top\mathbf{Q}\mathbf{\Sigma}^{1/2}$. With this, we have

$$\mathbf{D}\mathbf{G}^\top = (\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1} = \mathbf{I},$$

and

$$\widetilde{\mathbf{U}} = \mathbf{U}\mathbf{D} = \mathbf{P}\mathbf{\Sigma}^{1/2},\ \widetilde{\mathbf{V}} = \mathbf{V}\mathbf{G} = \mathbf{Q}\mathbf{\Sigma}^{1/2}.$$

The above constructed pair $(\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}})$ satisfies

$$\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top = \mathbf{U}\mathbf{V}^\top,\ \widetilde{\mathbf{U}}^\top\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}}^\top\widetilde{\mathbf{V}}.$$

Since $(\mathbf{U}, \mathbf{V}) \in \mathcal{C}$, we have

$$\nabla h_{\mathbf{U}}(\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}}) = \nabla h_{\mathbf{U}}(\mathbf{U}, \mathbf{V})\mathbf{D} = \mathbf{0},\ \nabla h_{\mathbf{V}}(\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}}) = \nabla h_{\mathbf{V}}(\mathbf{U}, \mathbf{V})\mathbf{G} = \mathbf{0},$$

which implies that $(\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}})$ is also a critical point (but not a global minimum since by assumption $(\mathbf{U}, \mathbf{V})$ is not a global minimum) of h. Since $(\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}})$ is also balanced, it follows from Lemma F.6.2 that there exists $\widetilde{\mathbf{\Delta}}_{\widetilde{\mathbf{U}}}$ and $\widetilde{\mathbf{\Delta}}_{\widetilde{\mathbf{V}}}$ such that

$$[\nabla^2 h(\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}})](\widetilde{\mathbf{\Delta}}, \widetilde{\mathbf{\Delta}}) < 0.$$

Now construct $\mathbf{\Delta_U} = \mathbf{\Delta}_{\widetilde{\mathbf{U}}}\mathbf{D}^{-1}$ and $\mathbf{\Delta_V} = \widetilde{\mathbf{\Delta}}_{\widetilde{\mathbf{V}}}\mathbf{G}^{-1}$. Then, we have

$$\begin{aligned}
[\nabla^2 h(\mathbf{U}, \mathbf{V})](\mathbf{\Delta}, \mathbf{\Delta}) &= \|\mathbf{\Delta_U}\mathbf{V}^\top + \mathbf{U}\mathbf{\Delta_V}^\top\|_F^2 + 2\langle \mathbf{U}\mathbf{V}^\top - \mathbf{Y}, \mathbf{\Delta_U}\mathbf{\Delta_V}^\top\rangle \\
&= \|\widetilde{\mathbf{\Delta}}_{\widetilde{\mathbf{U}}}\widetilde{\mathbf{V}}^\top + \widetilde{\mathbf{U}}\mathbf{\Delta}_{\widetilde{\mathbf{V}}}^\top\|_F^2 + 2\langle \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top - \mathbf{Y}, \mathbf{\Delta}_{\widetilde{\mathbf{U}}}\mathbf{\Delta}_{\widetilde{\mathbf{V}}}^\top\rangle \\
&= [\nabla^2 h(\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}})](\widetilde{\mathbf{\Delta}}, \widetilde{\mathbf{\Delta}}) < 0,
\end{aligned}$$

which implies that $(\mathbf{U}, \mathbf{V})$ is a strict saddle. □

Lemma F.6.2 together with Lemma F.6.3 implies that any pair $(\mathbf{U}, \mathbf{V}) \in \mathcal{C}$ is either a global minimum of h or a strict saddle.

□

APPENDIX G

APPENDICES FOR CHAPTER 10

G.1 Implementations and Numerical Experiments

G.1.1 Implementations: Closed-form Updating Formula

In this section, we discuss and give closed-form expressions for the iterates of Bregman gradient descent (Algorithm 3) and Bregman alternating gradient descent (Algorithm 4), respectively. Both Bregman proximal minimization (Algorithm 5) and Bregman proximal alternating minimization (Algorithm 6) are similar to standard (alternating) proximal minimization in that the existence of closed-form solutions depends on the specific form of the objective function f. Therefore, in this part we mainly focus on deriving closed-form expressions for Bregman gradient descent Algorithm 3 and Bregman alternating gradient descent (Algorithm 4), respectively.

For simplicity and generality, let us consider a fourth-degree polynomial objective function f.[47] This is because a fourth-degree polynomial objective function can cover a number of matrix factorization problems, such as matrix PCA, matrix sensing and matrix completion.

G.1.1.1 Closed-form Updating Formula for Bregman Gradient Decent

By Lemma 10.3.1, to obtain the second-order convergence of Bregman gradient descent for a fourth-degree polynomial objective function, it is sufficient to set the Bregman distance kernel $h(\mathbf{x})$ to be (i.e., using (10.22) with $d = 4$ and $\alpha = \sigma = 1$)

$$h(\mathbf{x}) = \frac{1}{4}\|\mathbf{x}\|_2^4 + \frac{1}{2}\|\mathbf{x}\|_2^2 + 1.$$

Now let us consider the main step (10.18) of Bregman gradient descent (Algorithm 3):

$$\mathbf{x}^\ell = \arg\min_{\mathbf{x}} f(\mathbf{x}^{\ell-1}) + \langle \nabla f(\mathbf{x}^{\ell-1}), \mathbf{x} - \mathbf{x}^{\ell-1}\rangle + \frac{1}{\eta} D_h(\mathbf{x}, \mathbf{x}^{\ell-1}).$$

Theorem G.1.1. *Suppose the objective function $f(\mathbf{x})$ is any fourth-degree polynomial. By Lemma 10.3.1, the Bregman distance kernel h can be set according to* (10.22) *with $d = 4$ and $\alpha = \sigma = 1$. Then there is a closed-form updating formula for the main step* (10.18) *of Bregman gradient descent (Algorithm 3) which is given by*

$$\mathbf{x}^\ell = \tau(\|\mathbf{z}^{\ell-1}\|_2^2)\mathbf{z}^{\ell-1} \tag{G.1}$$

where

[47] It is not difficult to consider an arbitrary dth- or (d_1, d_2)th-order of polynomial objective function. The only issue in this case is that there might be no closed-form updating formular (e.g., Theorem G.1.1) in solving the optimality condition of the updating formula (10.18) in Bregman gradient descent (Algorithm 3) or (10.19) in Bregman alternating gradient descent (Algorithm 4), but one can nevertheless solve the optimality condition using line-search algorithms.

$$\mathbf{z}^{\ell-1} := (\|\mathbf{x}^{\ell-1}\|_2^2 + 1)\mathbf{x}^{\ell-1} - \eta\nabla f(\mathbf{x}^{\ell-1})$$

and $\tau(\cdot)$ *is defined in* (G.4)*.*

Remark G.1.1. Therefore, we can view Bregman gradient descent (10.18) as standard gradient descent equipped with an adaptive choice of stepsize (see (G.1)).

Proof of Theorem G.1.1. First of all, the first-order optimality condition of (10.18) is given by letting gradient of the objective function of (10.18) vanish

$$\eta\nabla f(\mathbf{x}^{\ell-1}) - \nabla h(\mathbf{x}^{\ell-1}) + \nabla h(\mathbf{x}^{\ell}) = \mathbf{0},$$

which together with the fact $\nabla h(\mathbf{x}) = (\|\mathbf{x}\|_2^2 + 1)\mathbf{x}$ gives the new optimality condition:

$$(\|\mathbf{x}^{\ell}\|_2^2 + 1)\mathbf{x}^{\ell} = (\|\mathbf{x}^{\ell-1}\|_2^2 + 1)\mathbf{x}^{\ell-1} - \eta\nabla f(\mathbf{x}^{\ell-1}). \tag{G.2}$$

Thus, we can conclude that the closed-form update for Bregman gradient descent (Algorithm 3) is

$$\mathbf{x}^{\ell} = t_{\ell-1} \cdot [(\|\mathbf{x}^{\ell-1}\|_2^2 + 1)\mathbf{x}^{\ell-1} - \eta\nabla f(\mathbf{x}^{\ell-1})] := t_{\ell-1}\mathbf{z}^{\ell-1}, \tag{G.3}$$

where the scalar t_ℓ depends on the norm of current iterate

$$\mathbf{z}^{\ell-1} := (\|\mathbf{x}^{\ell-1}\|_2^2 + 1)\mathbf{x}^{\ell-1} - \eta\nabla f(\mathbf{x}^{\ell-1})$$

and is chosen so that the new optimality condition (G.2) is satisfied when we plug in (G.3). That is,

$$(\|t_{\ell-1}\mathbf{z}^{\ell-1}\|_2^2 + 1)(t_{\ell-1}\mathbf{z}^{\ell-1}) = \mathbf{z}^{\ell-1}.$$

By multiplying $\mathbf{z}^{\ell-1}$ on both sides, we get that $t_{\ell-1}$ should satisfy the following cubic polynomial equation

$$\|\mathbf{z}^{\ell-1}\|_2^2 t^3 + t - 1 = 0,$$

which can be shown to have a unique real solution with a closed-form expression (see Lemma G.1.1). Finally, combining Lemma G.1.1 and (G.3), we get the update step (10.18) in closed-form is

$$\mathbf{x}^{\ell} = \tau(\|\mathbf{z}^{\ell-1}\|_2^2)\mathbf{z}^{\ell-1},$$

where $\mathbf{z}^{\ell-1} := (\|\mathbf{x}^{\ell-1}\|_2^2 + 1)\mathbf{x}^{\ell-1} - \eta\nabla f(\mathbf{x}^{\ell-1})$ and $\tau(\cdot)$ is given in (G.4). □

Lemma G.1.1. *For any* $a \geq 0$*, the following cubic polynomial*

$$at^3 + t - 1 = 0$$

has a unique real solution depending on a*, that is,*

$$t = \tau(a) := \frac{\sqrt[3]{2}\left(9\sqrt{a} + \sqrt{3}\sqrt{27a+4}\right)^{2/3} - 2\sqrt[3]{3}}{6^{2/3}\sqrt[3]{\sqrt{3}\sqrt{a^3(27a+4)} + 9a^2}}, \tag{G.4}$$

where $\tau(a)$ is a decreasing function on $a > 0$ satisfying $\lim_{a\to 0^+} \tau(a) = 1$ and $\lim_{a\to\infty} \tau(a) = 0$.

Proof. Since $\phi_a'(t) = 3at^2 + 1 > 0$ for any $a \geq 0$ and any $t \in \mathbb{R}$, $\phi_a(t)$ is a strictly increasing function of t. Then noting that $\phi_a(0) = -1$ and $\phi_a(1) = a \geq 0$, we can conclude that $\phi_a(t) = 0$ has a unique real root lying within $(0, 1]$ in view of the continuity of $\phi_a(\cdot)$. Then by direct computations, this unique real root is given by

$$t = \tau(a) := \frac{\sqrt[3]{2}\left(9\sqrt{a} + \sqrt{3}\sqrt{27a+4}\right)^{2/3} - 2\sqrt[3]{3}}{6^{2/3}\sqrt[3]{\sqrt{3}\sqrt{a^3(27a+4)} + 9a^2}}.$$

The limit property is obtained by directly taking the limit of the function $\tau(a)$.

Now we show the decreasing property of the solution $\tau(a)$ with respect to a. This can be shown using the implicit function theorem that the following implicit equation holds

$$t^3 + 3at^2t'(a) + t'(a) = 0.$$

Directly solving this equation, we get

$$t'(a) = -\frac{t^3}{3at^2 + 1} < 0 \text{ for any } a \geq 0 \text{ and } t > 0.$$

By identifying that

$$t'(a) = \frac{d\tau(a)}{da},$$

we obtain that $\tau(a)$ is actually a decreasing function of a. □

G.1.1.2 Closed-form Updating Formula for Bregman alternating Gradient Decent

Now let us consider the case where the objective function $f(\mathbf{x}, \mathbf{y})$ is a $(4, 4)$th polynomial. Recall the main step (10.19) of Bregman alternating gradient descent (Algorithm 4):

$$\begin{aligned}
\mathbf{x}^\ell &= \arg\min_{\mathbf{x}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}), \mathbf{x} - \mathbf{x}^{\ell-1}\rangle + \frac{1}{\eta} D_h^1(\mathbf{x}, \mathbf{x}^{\ell-1}; \mathbf{y}^{\ell-1}), \\
\mathbf{y}^\ell &= \arg\min_{\mathbf{y}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}^{\ell}, \mathbf{y}^{\ell-1}), \mathbf{y} - \mathbf{y}^{\ell-1}\rangle + \frac{1}{\eta} D_h^2(\mathbf{y}, \mathbf{y}^{\ell-1}; \mathbf{x}^{\ell}).
\end{aligned}$$

By recognizing the similar structure in the main step (10.18) of Bregman gradient descent Algorithm 3 and the main step (10.19) of Bregman alternating gradient descent Algorithm 4, we can easily use Theorem G.1.1 to derive the closed-form updating formula for the Bregman alternating gradient descent Algorithm 4 where the objective function $f(\mathbf{x}, \mathbf{y})$ is a $(4, 4)$th polynomial.

Corollary G.1.1. *Suppose the objective function $f(\mathbf{x}, \mathbf{y})$ is any $(4, 4)$th-degree polynomial. By Lemma 10.3.1, the bi-Bregman distance kernel $h(\mathbf{x}, \mathbf{y})$ as*

$$h(\mathbf{x}, \mathbf{y}) = \left(\frac{1}{4}\|\mathbf{x}\|_2^4 + \frac{1}{2}\|\mathbf{x}\|_2^2 + 1\right)\left(\frac{1}{4}\|\mathbf{y}\|_2^4 + \frac{1}{2}\|\mathbf{y}\|_2^2 + 1\right).$$

Then there is a closed-form updating formula for the main step (10.19) *of Bregman alternating gradient descent Algorithm 4 which is given by*

$$\begin{aligned} \mathbf{x}^\ell =& \tau(\|\mathbf{z}_x^{\ell-1}\|_2^2)\mathbf{z}_x^{\ell-1}, \\ \mathbf{y}^\ell =& \tau(\|\mathbf{z}_y^{\ell-1}\|_2^2)\mathbf{z}_y^{\ell-1}, \end{aligned} \tag{G.5}$$

with

$$\begin{aligned} \mathbf{z}_x^{\ell-1} &:= (\|\mathbf{x}^{\ell-1}\|_2^2 + 1)\mathbf{x}^{\ell-1} - \frac{\eta}{(\|\mathbf{y}^{\ell-1}\|_2^4/4 + \|\mathbf{y}^{\ell-1}\|_2^2/2 + 1)}\nabla_{\mathbf{x}} f(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}), \\ \mathbf{z}_y^{\ell-1} &:= (\|\mathbf{y}^{\ell-1}\|_2^2 + 1)\mathbf{y}^{\ell-1} - \frac{\eta}{(\|\mathbf{x}^{\ell}\|_2^4/4 + \|\mathbf{x}^{\ell}\|_2^2/2 + 1)}\nabla_{\mathbf{y}} f(\mathbf{x}^{\ell}, \mathbf{y}^{\ell-1}), \end{aligned}$$

and $\tau(\cdot)$ defined in (G.4)*.*

Proof. The proof of Corollary G.1.1 is similar to that of Theorem G.1.1. □

When the objective function $f(\mathbf{x}, \mathbf{y})$ is a $(2, 2)$th-degree polynomial, we can further simplify the closed-form updating formula (G.5) in Corollary G.1.1. We note that the importance of the $(2, 2)$th-degree polynomial objective functions comes from that it can also cover a massive number of matrix factorization problems, such as nonsymmetric matrix PCA, nonsymmetric matrix sensing and nonsymmetric matrix completion.

Theorem G.1.2. *Suppose the objective function $f(\mathbf{x}, \mathbf{y})$ is any $(2, 2)$th-degree polynomial. By Lemma 10.3.1, we can set the bi-Bregman distance kernel $h(\mathbf{x}, \mathbf{y})$ as*

$$h(\mathbf{x}, \mathbf{y}) = \left(\frac{1}{2}\|\mathbf{x}\|_2^2 + 1\right)\left(\frac{1}{2}\|\mathbf{y}\|_2^2 + 1\right). \tag{G.6}$$

Then there is a closed-form updating formula for the main step (10.19) *of Bregman alternating gradient descent Algorithm 4 which is given by*

$$\begin{aligned} \mathbf{x}^\ell =& \mathbf{x}^{\ell-1} - \frac{\eta}{\|\mathbf{y}^{\ell-1}\|_2^2/2 + 1}\nabla_{\mathbf{x}} f(\mathbf{x}, ^{\ell-1}, \mathbf{y}^{\ell-1}), \\ \mathbf{y}^\ell =& \mathbf{y}^{\ell-1} - \frac{\eta}{\|\mathbf{x}^{\ell}\|_2^2/2 + 1}\nabla_{\mathbf{y}} f(\mathbf{x}^{\ell}, \mathbf{y}^{\ell-1}). \end{aligned} \tag{G.7}$$

Remark G.1.2. In view of the closed-form updating formula (G.7), Bregman alternating gradient descent can be viewed as the standard proximal alternating linearized minimization (a.k.a. alternating gradient descent) equipped with the ability of adaptively choosing the proximal regularizer parameter.

Proof of Theorem G.1.2. Recall the main step (10.19) of Bregman alternating gradient descent (Algorithm 4):

$$\begin{aligned}\mathbf{x}^\ell &= \arg\min_{\mathbf{x}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}), \mathbf{x} - \mathbf{x}^{\ell-1}\rangle + \frac{1}{\eta} D_h^1(\mathbf{x}, \mathbf{x}^{\ell-1}; \mathbf{y}^{\ell-1}),\\ \mathbf{y}^\ell &= \arg\min_{\mathbf{y}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}^{\ell}, \mathbf{y}^{\ell-1}), \mathbf{y} - \mathbf{y}^{\ell-1}\rangle + \frac{1}{\eta} D_h^2(\mathbf{y}, \mathbf{y}^{\ell-1}; \mathbf{x}^{\ell}).\end{aligned} \tag{G.8}$$

and the first and second Bregman distances are respectively defined as as (see Definition 10.2.3):

$$\begin{aligned}D_h^1(\mathbf{x}, \mathbf{x}^{\ell-1}; \mathbf{y}^{\ell-1}) =& h(\mathbf{x}, \mathbf{y}^{\ell-1}) - h(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}) - \langle \nabla_{\mathbf{x}} h(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}), \mathbf{x} - \mathbf{x}^{\ell-1}\rangle,\\ D_h^2(\mathbf{y}, \mathbf{y}^{\ell-1}; \mathbf{x}^{\ell}) =& h(\mathbf{x}^{\ell}, \mathbf{y}) - h(\mathbf{x}^{\ell}, \mathbf{y}^{\ell-1}) - \langle \nabla_{\mathbf{y}} h(\mathbf{x}^{\ell}, \mathbf{y}^{\ell-1}), \mathbf{y} - \mathbf{y}^{\ell-1}\rangle\end{aligned} \tag{G.9}$$

Now combining (G.8) and (G.10), we get the optimality conditions that should be satisfied by $(\mathbf{x}^\ell, \mathbf{y}^\ell)$:

$$\begin{aligned}\eta \nabla_{\mathbf{x}} f(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}) + \nabla_{\mathbf{x}} h(\mathbf{x}^{\ell}, \mathbf{y}^{\ell-1}) - \nabla_{\mathbf{x}} h(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}) = 0;\\ \eta \nabla_{\mathbf{y}} f(\mathbf{x}^{\ell}, \mathbf{y}^{\ell}) + \nabla_{\mathbf{y}} h(\mathbf{x}^{\ell}, \mathbf{y}^{\ell}) - \nabla_{\mathbf{x}} h(\mathbf{x}^{\ell}, \mathbf{y}^{\ell-1}) = 0;\end{aligned} \tag{G.10}$$

Finally, we obtain the closed-form updating formula (G.7) by combining the optimality conditions (G.10) and

$$\begin{aligned}\nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{y}) &= \left(\frac{1}{2}\|\mathbf{y}\|_2^2 + 1\right)\mathbf{x},\\ \nabla_{\mathbf{y}} h(\mathbf{x}, \mathbf{y}) &= \left(\frac{1}{2}\|\mathbf{x}\|_2^2 + 1\right)\mathbf{y}.\end{aligned}$$

□

G.1.2 Numerical Experiments on Low-rank Matrix Factorization

G.1.2.1 Low-rank Matrix Factorization Problem

As we have discussed in Section 10.3.3, the Burer-Monteiro factorization method [119,159] transforms the original large-scale rank-constrained matrix optimization problem (10.23):

$$\underset{\mathbf{X}\in\mathbb{S}_+^n \text{ or } \mathbf{X}\in\mathbb{R}^{n\times m}}{\text{minimize}} \; q(\mathbf{X}) \;\; \text{subject to} \;\; \operatorname{rank}(\mathbf{X}) \le r,$$

into a smaller-scale (nonconvex) problem (10.24):

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\text{minimize}}\, f(\mathbf{U}) := q(\mathbf{U}\mathbf{U}^\top) \quad \text{or} \quad \underset{\mathbf{U}\in\mathbb{R}^{n\times r}, \mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}}\, f(\mathbf{U}, \mathbf{V}) := q(\mathbf{U}\mathbf{V}^\top).$$

From Corollary 10.3.3, when the original objective function $q(\mathbf{X})$ in (10.23) is any lower-bounded dth or (d_1, d_2)th-degree polynomial satisfying the $(2r, \frac{1}{20})$-RIP and we set

$$h(\mathbf{U}) = \frac{\alpha}{d}\|\mathbf{U}\|_F^d + \frac{\sigma}{2}\|\mathbf{U}\|_F^2 + 1 \;\; \text{or} \;\; h(\mathbf{U}, \mathbf{V}) = \left(\frac{\alpha}{d_1}\|\mathbf{U}\|_F^{d_1} + \frac{\sigma}{2}\|\mathbf{U}\|_F^2 + 1\right)\left(\frac{\alpha}{d_2}\|\mathbf{V}\|_F^{d_2} + \frac{\sigma}{2}\|\mathbf{V}\|_F^2 + 1\right) \tag{G.11}$$

for any $\alpha, \sigma > 0$, then applying Algorithms 3 and 5 to $\text{minimize}_{\mathbf{U}}\, f(\mathbf{U})$ in (10.24) or applying Algorithms 4 and 6 to $\text{minimize}_{\mathbf{U},\mathbf{V}}\, f(\mathbf{U}, \mathbf{V})$ in (10.24), we can almost surely solve (10.23) to global optimality from a random initializa-

tion.

The global optimality results (i.e., Corollary 10.3.3) hold for any lower-bounded finite-degree[48] polynomial objective function $q(\mathbf{X})$ satisfying $(2r, \frac{1}{20})$-RIP. Now for convenience, let us consider the best rank-r approximation problem of the optimization form

$$\underset{\mathbf{X}\in\mathbb{S}^n_+ \text{ or } \mathbf{X}\in\mathbb{R}^{n\times m}}{\text{minimize}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}^\star\|_F^2 \ \text{ subject to } \ \text{rank}(\mathbf{X}) \le r. \tag{G.12}$$

In this case of (G.12), we know that the original objective function $q(\mathbf{X})$ satisfies $(2r, \delta)$-RIP for any positive integer r and any $\delta \ge 0$ by recognizing that $\nabla^2 q(\mathbf{X}) = \mathbf{I}$. Therefore, the global optimality theory (i.e., Corollary 10.3.3) can directly apply to this case.

Now recall that the motivation behind the BM factorization comes from the high computational/storage cost of solving large-scale matrix optimization problems, particularly those involving rank constraints. We therefore are more interested in the BM formulation of the rank-r approximation problem (which is the *matrix factorization* problem), that is

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}}{\text{minimize}}\, f(\mathbf{U}) := \frac{1}{2}\|\mathbf{U}\mathbf{U}^\top-\mathbf{X}^\star\|_F^2 \quad \text{or} \quad \underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \frac{1}{2}\|\mathbf{U}\mathbf{V}^\top-\mathbf{X}^\star\|_F^2. \tag{G.13}$$

G.1.2.2 Implementations and Experiments

In (G.13), we identify that $f(\mathbf{U})$ is a fourth-degree polynomial and $f(\mathbf{U},\mathbf{V})$ is a $(2,2)$th-degree polynomial. Therefore, we can use the closed-form updating formulas given in Theorem G.1.1 and Theorem G.1.2 to solve $\text{minimize}_{\mathbf{U}}\, f(\mathbf{U})$ (for the symmetric case) and $\text{minimize}_{\mathbf{U},\mathbf{V}}\, f(\mathbf{U},\mathbf{V})$ (for the symmetric case), respectively. More precisely,

- we use the closed-form updating formula (i.e., (G.1) with $\mathbf{x}^\ell = \mathbf{U}^\ell$ (for the symmetric case) to perform Bregman gradient descent (Algorithm 3) on the symmetric matrix factorization problem in (G.13), and
- use the closed-form updating formula (G.7) with $(\mathbf{x}^\ell, \mathbf{y}^\ell) = (\mathbf{U}^\ell, \mathbf{V}^\ell)$ for the nonsymmetric case to perform Bregman alternating gradient descent (Algorithm 4) on the nonsymmetric matrix factorization problem in (G.13).
- Moreover, to verify the global optimality theory of Algorithms 3 and 4 in solving Burer-Monteiro factorization problems (G.1.1) (see Corollary 10.3.3), we will plot the optimality distances $\|\mathbf{U}^\ell\mathbf{U}^{\ell\top}-\mathbf{X}_r^\star\|_F^2$ (for symmetric case) and $\|\mathbf{U}^\ell\mathbf{V}^{\ell\top}-\mathbf{X}_r^\star\|_F^2$ (for nonsymmetric case) as a function of the number of iterations. This is because all the second-order stationary points of (G.13) correspond to the best rank-r approximation of $\mathbf{X}^\star$, denoted by $\mathbf{X}_r^\star$ (cf. [6, 206]).

[48] Clearly, the finite-degree of the polynomial $q(\mathbf{X})$ directly implies the finite-degree of the polynomial $f(\mathbf{U})$ or $f(\mathbf{U},\mathbf{V})$.

Now we compare the performances of the standard (alternating) gradient descent and the Bregman (alternating) gradient descent in solving the symmetric and nonsymmetric matrix factorization problems in (G.13). For setting up the parameters, we set the dimensions as $n = m = 50$, rank as $r = 2$, and for convenience we generate the ground truth matrix $\mathbf{X}^\star$ with $\mathrm{rank}(\mathbf{X}^\star) = 5$. We define $\mathbf{X}_r^\star$ as the best rank-r approximation of the ground truth $\mathbf{X}^\star$, which can be obtained via SVD. In term of the choices of the stepsizes (a.k.a. regularizer parameters) η, we remark that we manually tune the stepsizes η to achieve the best performances for all the algorithms. In term of standard (alternating) gradient descent, because by (10.5), to ensure the sufficient decrease property, the stepsize should be less than $\frac{2}{L_f}$, where L_f is the global Lipschitz gradient constant for the objective function to be minimized. However, for the matrix factorization problem, there is no global Lipschitz constant. Therefore, tune the stepsize for each different initialization. In sharp contrast to standard (alternating) gradient descent, Bregman (alternating) gradient descent is not sensitive to the choice of the "stepsize" η. This is because, as described in Remark G.1.1 and Remark G.1.2, Bregman (alternating) gradient descent uses a kind of adaptive stepsize according to the current iteration $\mathbf{U}^k$ (see (G.1)). To illustrate of such ability of adaptively-choosing stepsize, we generate two random initialization points: one is close to the origin and the other is far away from the origin (and hence has large norms). Figure Figure G.1 shows that when initialized at a point having a large norm, the performance of (alternating) gradient descent degrades drastically, while Bregman (alternating) gradient descent maintains a stable and favorable performance regardless of the size of the initialization.

G.1.2.3 More Experiments for Algorithm 6

Finally, we are interested in deriving a closed-form expression of Bregman proximal alternating minimization (Algorithm 6) for the nonsymmetric matrix factorization problem:

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\|_F^2. \tag{G.14}$$

Theorem G.1.3. *Suppose $f(\mathbf{U},\mathbf{V})$ is the objective function of* (G.14) *. Set the bi-Bregman kernel $h(\mathbf{U},\mathbf{V})$ according to* (G.6)*. Then the closed-form updating formula of Bregman proximal alternating minimization (Algorithm 6) is given by*

$$\begin{aligned}
\mathbf{U}^\ell &= \left(\frac{\eta}{\|\mathbf{V}^{\ell-1}\|_F^2/2+1}\mathbf{X}^\star\mathbf{V}^{\ell-1} + \mathbf{U}^{\ell-1}\right)\left(\frac{\eta}{\|\mathbf{V}^{\ell-1}\|_F^2/2+1}\mathbf{V}^{\ell-1\top}\mathbf{V}^{\ell-1} + \mathbf{I}\right)^{-1} \\
\mathbf{V}^\ell &= \left(\frac{\eta}{\|\mathbf{U}^{\ell}\|_F^2/2+1}\mathbf{X}^{\star\top}\mathbf{U}^{\ell} + \mathbf{V}^{\ell-1}\right)\left(\frac{\eta}{\|\mathbf{U}^{\ell}\|_F^2/2+1}\mathbf{U}^{\ell\top}\mathbf{U}^{\ell} + \mathbf{I}\right)^{-1}
\end{aligned} \tag{G.15}$$

Proof. The proof directly can be obtained by combining the optimality conditions of the main step (10.21) of Algorithm 6 (with $f(\mathbf{U},\mathbf{V})$ set as (G.14)) and

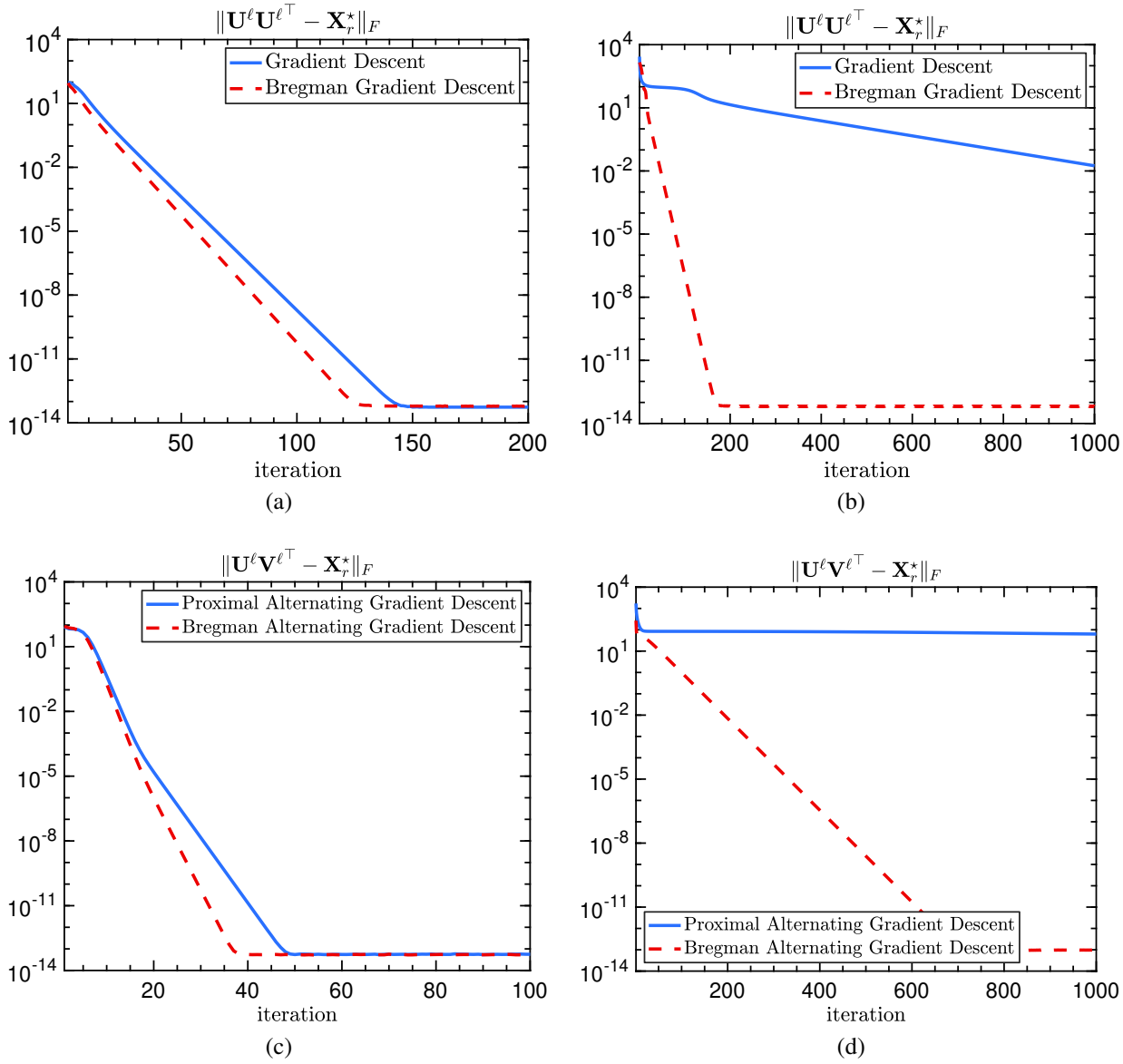


Figure G.1: Comparing standard (alternating) gradient descent and Bregman (alternating) gradient descent in solving symmetric and nonsymmetric matrix factorizations in (G.13). In particular, we set up the symmetric matrix factorization experiments as follows. (a): We initialize $\mathbf{U}^0$ with each entry drawn from $\mathcal{N}(0,1)$; (b): We initialize $\mathbf{U}^0$ with each entry drawn from $\mathcal{N}(0,100)$. We note that in both cases (a) and (b), we have tuned the stepsizes of both algorithms to achieve optimal performance. We observe that when the current $\|\mathbf{U}\|_F$ is large, the convergence of gradient descent becomes very slow; while Bregman gradient descent is not sensitive to the norm of the current $\|\mathbf{U}\|_F$ and still converges quickly to the global optimum. The same phenomenon happens in non-symmetric matrix factorization. (c): We initialize $\mathbf{U}^0$ and $\mathbf{V}^0$ with each entry drawn from $\mathcal{N}(0,1)$; (d): We initialize $\mathbf{U}^0$ and $\mathbf{V}^0$ with each entry drawn from $\mathcal{N}(0,100)$. Similar to the symmetric case, we have tuned the stepsize of both algorithms to achieve optimal performance in both cases. We observe that the performance of (alternating) gradient descent degrades drastically and even fails (see (d)), while Bregman (alternating) gradient descent maintains a stable and favorable performance regardless of the size of the initialization.

$$\nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{y}) = \left(\frac{1}{2}\|\mathbf{y}\|_2^2 + 1\right)\mathbf{x},$$
$$\nabla_{\mathbf{y}} h(\mathbf{x}, \mathbf{y}) = \left(\frac{1}{2}\|\mathbf{x}\|_2^2 + 1\right)\mathbf{y}.$$

□

Remark G.1.3. Now we can compare the closed-form updating formula (G.15) of Bregma proximal alternating minimization with that of standard proximal alternating minimization in solving (G.14), which is given by

$$\begin{aligned}\mathbf{U}^{\ell} &= \left(\eta \mathbf{X}^{\star}\mathbf{V}^{\ell-1} + \mathbf{U}^{\ell-1}\right)\left(\eta \mathbf{V}^{\ell-1\top}\mathbf{V}^{\ell-1} + \mathbf{I}\right)^{-1} \\ \mathbf{V}^{\ell} &= \left(\eta \mathbf{X}^{\star\top}\mathbf{U}^{\ell} + \mathbf{V}^{\ell-1}\right)\left(\eta \mathbf{U}^{\ell\top}\mathbf{U}^{\ell} + \mathbf{I}\right)^{-1}\end{aligned} \tag{G.16}$$

Therefore, we can view the closed-form updating formula (G.15) of Bregman proximal alternating minimization as that of standard proximal alternating minimization equipped with adaptive stepsize according to the norm of the current iteration $\mathbf{V}^{\ell-1}$ or $\mathbf{U}^{\ell}$.

We will compare the performances of standard proximal alternating minimization and Bregman proximal alternating minimization in solving the nonsymmetric matrix factorization problem (G.14). For convenience, we set up the experiments exactly the same as that of the the first experiment (see Figure Figure G.1) except for those closed-form updating formulas. To verify our convergence theories, we plot the optimality distance $\|\mathbf{U}^{\ell}\mathbf{V}^{\ell\top} - \mathbf{X}_r^{\star}\|_F^2$ as a function of the number of iterations. In the experiments (see Figure Figure G.2), we observe that both standard and Bregman proximal alternating minimizations achieve amazingly satisfying performance in solving the nonsymmetric matrix factorization problem (G.14). Further, both algorithms can maintain a stable and favorable performance regardless of the size of the initialization. In our understanding, this findings can be explained by Equation (10.17) (similar results hold for standard proximal alternating minimization if we replace the L_f-adaptive-Lipschitz gradient condition by the standard L_f-Lipschitz gradient condition), which essentially claims that the amount of sufficient decrease is independent of the (adaptive) Lipschitz constant L_f, hence allowing more flexibility in tuning η.

G.2 Proof of Lemma 10.2.1

Lemma G.2.1 (Lemma 10.2.1)**.** *Suppose $f \in \mathcal{C}^2$ is globally lower bounded and satisfies the L_f-adaptive Lipschitz gradient condition for some Bregman distance kernel $h \in \mathcal{C}^2$, which is assumed to be σ-strongly convex and supercoercive. Then the updating formula* (10.18) *for Bregman gradient descent (Algorithm 5) and* (10.20) *for Bregman proximal minimization (Algorithm 5) are both well-defined and respectively satisfy:*

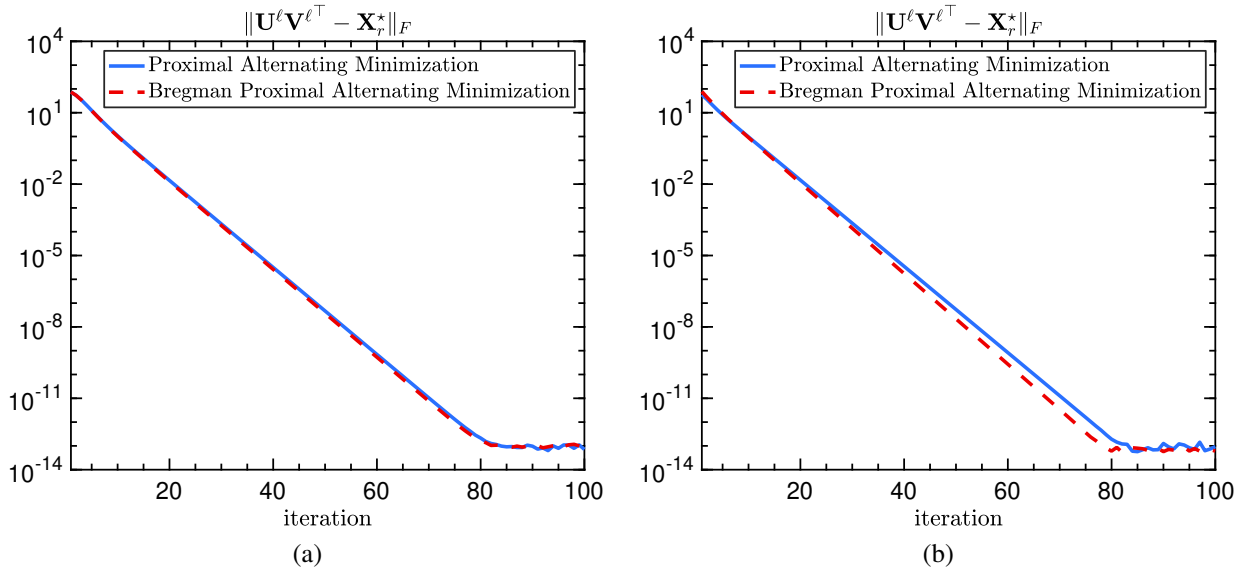


Figure G.2: Comparing standard proximal alternating minimization and Bregman proximal alternating minimization in solving the nonsymmetric matrix factorization problem (G.14). In particular, we set up experiments as follows. (a): We initialize $\mathbf{U}^0$ and $\mathbf{V}^0$ with each entry drawn from $\mathcal{N}(0,1)$; (b): We initialize $\mathbf{U}^0$ and $\mathbf{V}^0$ with each entry drawn from $\mathcal{N}(0,100)$. We note that in both cases, we have tuned the proximal regularization parameter η for both standard and Bregman proximal alternating minimization algorithms to achieve optimal performance. We observe that both algorithms can maintain a stable and favorable performance regardless of the size of the initialization.

$$\text{Algorithm 3}: \quad f(\mathbf{x}^{\ell-1}) - f(\mathbf{x}^\ell) \geq \left(\frac{1}{\eta} - L_f\right)\frac{\sigma}{2}\|\mathbf{x}^\ell - \mathbf{x}^{\ell-1}\|_2^2 \tag{10.10}$$

$$\text{Algorithm 5}: \quad f(\mathbf{x}^{\ell-1}) - f(\mathbf{x}^\ell) \geq \frac{\sigma}{2\eta}\|\mathbf{x}^\ell - \mathbf{x}^{\ell-1}\|_2^2. \tag{10.11}$$

Proof. We first show (10.10). For simplifying notations, denote $\mathbf{x}^+ := \mathbf{x}^\ell$ and $\mathbf{x}^{\ell-1} := \mathbf{x}^-$. For the well-definedness, it suffices to show the solution of (10.10) exists and is unique. First, since the objective function is continuous (as $f, h \in \mathcal{C}^2$), its level set

$$\text{Lev}_\phi(a) := \{\mathbf{x} : \phi(\mathbf{x}) \leq a\}$$

is closed for any $a \in \mathbb{R}$, where $\phi(\mathbf{x}) := f(\mathbf{x}^-) + \langle \nabla f(\mathbf{x}^-), \mathbf{x} - \mathbf{x}^-\rangle + \frac{1}{\eta} D_h(\mathbf{x}, \mathbf{x}^-)$. Second, when h is super-coercive, we will show the objective function $\phi(\mathbf{x})$ is coercive, which would imply the boundedness of the level set $\text{Lev}_\phi(a)$. Then combing the closedness of the level set, we can view (10.10) as a minimization of a continuous function over a compact level set and hence the solution must exist. The uniqueness follows from the strong convexity of ϕ because $\nabla^2\phi = \nabla^2 h$ and h is strongly convex. Now, we show ϕ is coercive.

$$
\begin{aligned}
\phi(\mathbf{x}) =& f(\mathbf{x}^-) + \langle \nabla f(\mathbf{x}^-), \mathbf{x} - \mathbf{x}^- \rangle + \frac{1}{\eta} D_h(\mathbf{x}, \mathbf{x}^-) \\
=& f(\mathbf{x}^-) + \langle \nabla f(\mathbf{x}^-), \mathbf{x} - \mathbf{x}^- \rangle + \frac{1}{\eta} (h(\mathbf{x}) - h(\mathbf{x}^-) - \langle \nabla h(\mathbf{x}^-), \mathbf{x} - \mathbf{x}^- \rangle) \\
=& \frac{h(\mathbf{x})}{\eta} + \left\langle \nabla f(\mathbf{x}^-) - \frac{\nabla h(\mathbf{x}^-)}{\eta}, \mathbf{x} \right\rangle + \left(f(\mathbf{x}^-) - \left\langle \nabla f(\mathbf{x}^-), \mathbf{x}^- \right\rangle - \frac{h(\mathbf{x}^-)}{\eta} + \left\langle \frac{\nabla h(\mathbf{x}^-)}{\eta}, \mathbf{x}^- \right\rangle \right) \\
:=& \frac{h(\mathbf{x})}{\eta} + \langle \mathbf{a}, \mathbf{x} \rangle + b \\
=& \|\mathbf{x}\|_2 \left(\frac{h(\mathbf{x})}{\eta \|\mathbf{x}\|_2} + \left\langle \mathbf{a}, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \right\rangle \right) + b
\end{aligned}
$$

Therefore, we have

$$
\phi(\mathbf{x}) \geq \|\mathbf{x}\|_2 \left(\frac{h(\mathbf{x})}{\eta \|\mathbf{x}\|_2} - \|\mathbf{a}\|_2 \right) + b
$$

Together with that $h(\mathbf{x})/\|\mathbf{x}\|_2 > \eta \|\mathbf{a}\|_2$ for some large enough $\|\mathbf{x}\|_2$ by the super-coercivity of h, this shows that ϕ is coercive.

For the sufficient decrease property of (10.10), using definition of global optimality, we have

$$
\begin{aligned}
f(\mathbf{x}^-) &= f(\mathbf{x}^-) + \langle \nabla f(\mathbf{x}^-), \mathbf{x} - \mathbf{x}^- \rangle + \frac{1}{\eta} D_h(\mathbf{x}, \mathbf{x}^-) \Big|_{\mathbf{x}=\mathbf{x}^-} \\
&\geq f(\mathbf{x}^-) + \langle \nabla f(\mathbf{x}^-), \mathbf{x}^+ - \mathbf{x}^- \rangle + \frac{1}{\eta} D_h(\mathbf{x}^+, \mathbf{x}^-) \\
&\geq f(\mathbf{x}^+) - L_f D_h(\mathbf{x}^+, \mathbf{x}^-) + \frac{1}{\eta} D_h(\mathbf{x}^+, \mathbf{x}^-) \\
&= f(\mathbf{x}^+) + \left(\frac{1}{\eta} - L_f \right) D_h(\mathbf{x}^+, \mathbf{x}^-) \\
&\geq f(\mathbf{x}^+) + \left(\frac{1}{\eta} - L_f \right) \frac{\sigma}{2} \|\mathbf{x}^+ - \mathbf{x}^-\|_2^2
\end{aligned}
\tag{G.17}
$$

where the second inequality is by the general descent lemma (10.9) with $\mathbf{y} = \mathbf{x}^-, \mathbf{x} = \mathbf{x}^+$ and the last inequality follows from the σ-strong convexity of h.

We now show (10.11). Its well-definedness follows in the same way by showing that the objective function of (10.11) is coercive (by using the same analysis as (G.17) combined with the lower-boundedness of f) and strongly convex (since (f, h) satifies L_f-adaptive Lipschitz gradient condition and $\eta \in (0, 1/L_f)$). And the sufficient decrease property follows from the definition of global optimality,

$$
f(\mathbf{x}^-) = f(\mathbf{x}) + \frac{1}{\eta} D_h(\mathbf{x}, \mathbf{x}^-) \Big|_{\mathbf{x}=\mathbf{x}^-} \geq f(\mathbf{x}^+) + \frac{1}{\eta} D_h(\mathbf{x}^+, \mathbf{x}^-) \geq f(\mathbf{x}^+) + \frac{\sigma}{2\eta} \|\mathbf{x}^+ - \mathbf{x}^-\|_2^2.
$$

□

G.3 Proofs in Section of Stylized Applications

In this section, we collect the proofs omitted in the section of Stylized Applications.

G.3.1 Application to Polynomial Objective Functions

Lemma G.3.1 (Lemma 10.3.1)**.** *Suppose $f(\mathbf{x})$ (or $f(\mathbf{x},\mathbf{y})$) is any coercive and lower-bounded dth-order (or (d_1,d_2)th-order) polynomial function with $d, d_1, d_2 \geq 2$. Set the Bregman (or bi-Bregman) distance kernel h to be*

$$h(\mathbf{x}) = \frac{\alpha}{d}\|\mathbf{x}\|_2^d + \frac{\sigma}{2}\|\mathbf{x}\|_2^2 + 1 \quad \text{or} \quad h(\mathbf{x},\mathbf{y}) = (\frac{\alpha}{d_1}\|\mathbf{x}\|_2^{d_1} + \frac{\sigma}{2}\|\mathbf{x}\|_2^2 + 1)(\frac{\alpha}{d_2}\|\mathbf{y}\|_2^{d_2} + \frac{\sigma}{2}\|\mathbf{y}\|_2^2 + 1) \tag{G.18}$$

for any $\alpha, \sigma > 0$. Then $(f(\mathbf{x}), h(\mathbf{x}))$ (or $(f(\mathbf{x},\mathbf{y}), h(\mathbf{x},\mathbf{y}))$) satisfies Assumptions 10.2.1– 10.2.4.

Proof. We classify the proof into two parts according to two different cases of (f, h).

Showing $(f(\mathbf{x}), h(\mathbf{x}))$ satisfying Assumptions 10.2.1- 10.2.4. First, any dth-degree polynomial function $f(\mathbf{x})$ can be represented as

$$f(\mathbf{x}) = \sum_{k=0}^{d} \langle \mathcal{A}_k, \mathbf{x}^{\otimes k} \rangle \tag{G.19}$$

where $\otimes$ denotes the tensor product operator, $\mathbf{x}^{\otimes k} := \mathbf{x}\otimes\mathbf{x}\otimes\cdots\otimes\mathbf{x}$ (total k times), and the coefficients of kth-degree monomials are arranged as $\mathcal{A}_k \in \mathbb{R}^n \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ (total k times). For convenience, we denote $\mathbf{x}^{\otimes 0} = 1$ and $\mathcal{A}_0 \in \mathbb{R}$. Further, due to super-symmetric tensors $\mathbf{x}^{\otimes k}$ (i.e., the $(i_1, i_2, \cdots, i_k)$th entry of $\mathbf{x}^{\otimes k}$ is invariant respect to the order the indices $i_1, i_2, \ldots, i_k$), we can assume $\mathcal{A}_k$ for $k \geq 2$ as super-symmetric tensors since otherwise we can replace $\mathcal{A}$ as its super-symmetric part. For example, when $k = 2$ (i.e., $\mathcal{A}_2$ is a square matrix), the super-symmetric part for $\mathcal{A}_2$ is $(\mathcal{A}_2 + \mathcal{A}_2^\top)/2$. Similar definitions can be easily extending to a general natural number $k \geq 2$.

Now, we show that polynomial f with the particular Bregman distance kernel $h(\mathbf{x}) = \frac{\alpha}{d}\|\mathbf{x}\|_2^d + \frac{\beta}{2}\|\mathbf{x}\|_2^2 + 1$ in (10.22) satisfies all Assumptions 10.2.1- 10.2.4.

Showing Assumption 10.2.1. First of all, let us compute the Hessian of $h(\mathbf{x})$:

$$\begin{aligned}\nabla^2 h(\mathbf{x}) &= \frac{\alpha}{d} d(d-2)\|\mathbf{x}\|_2^{d-4}\mathbf{x}\mathbf{x}^\top + \left(\frac{\alpha}{d} d\|\mathbf{x}\|_2^{d-2} + 2\frac{\sigma}{2}\right)\mathbf{I}_n \\ &= \alpha(d-2)\|\mathbf{x}\|_2^{d-2}\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\frac{\mathbf{x}^\top}{\|\mathbf{x}\|_2} + \left(\alpha\|\mathbf{x}\|_2^{d-2} + \sigma\right)\mathbf{I}_n \end{aligned} \tag{G.20}$$

Therefore, we have $\nabla^2 h(\mathbf{x})$ is well-defined in the whole domain $\mathbf{x} \in \mathbb{R}^n$ for all $d \geq 2$, implying that $h \in \mathcal{C}^2$. Second,

$$\lim_{\|\mathbf{x}\|_2\to\infty} \frac{h(\mathbf{x})}{\|\mathbf{x}\|_2} = \lim_{\|\mathbf{x}\|_2\to\infty} \left(\frac{\alpha}{d}\|\mathbf{x}\|_2^{d-1} + \frac{\sigma}{2}\|\mathbf{x}\|_2 + \frac{1}{\|\mathbf{x}\|_2}\right) \geq \lim_{\|\mathbf{x}\|_2\to\infty} \frac{\sigma}{2}\|\mathbf{x}\|_2 = \infty,$$

implying the super-coercivity of h. Finally, following from (G.20), we can lower bound the Hessian as

$$\nabla^2 h(\mathbf{x}) \succeq \left(\alpha\|\mathbf{x}\|_2^{d-2} + \sigma\right)\mathbf{I}_n \succeq \sigma\mathbf{I}_n \tag{G.21}$$

indicating that $h(\mathbf{x})$ is σ-strong convex. This completes the proof of showing Assumption 10.2.1 holds.

Showing Assumption 10.2.2. First, the lower-boundedness of f follows by the assumption. Second, Since any polynomial function satisfies the KL property, f is a KL function by Remark 10.2.1. This completes the proof of showing Assumption 10.2.2 holds.

Showing Assumption 10.2.3. By definition of L_f-adaptive Lipschitz gradient condition, it suffices to show that there exists a constant L_f such that

$$L_f\nabla^2 h(\mathbf{x}) \pm \nabla^2 f(\mathbf{x}) \succeq 0 \quad \text{for all } \mathbf{x}$$

Towards that end, we first bound the Hessian spectral norm of f. Since f is a dth-degree polynomial function (i.e., in the form (G.19)), we can compute its Hessian matrix as (using the super-symmetry of $\mathcal{A}_k$)

$$\nabla^2 f(\mathbf{x}) = \sum_{k=2}^{d} k(k-1)\mathcal{A}_k \times_1 \mathbf{x} \times_2 \mathbf{x} \times_3 \mathbf{x} \cdots \times_{k-2} \mathbf{x} \tag{G.22}$$

where $\times_k$ denotes the kth-mode tensor-vector product [76] for any Nth-order tensor $\mathcal{A}$ and any vector $\mathbf{x} \in \mathbb{R}^n$ so that

$$\mathcal{A} \times_k \mathbf{x} = \left[\sum_{j=1}^{n} \mathcal{A}(i_1, \cdots, i_{k-1}, j, i_{k+1}, \cdots, i_N)\mathbf{x}(j)\right]_{i_1,\cdots,i_{k-1},i_{k+1},\cdots,i_N} \tag{G.23}$$

Using the triangle inequality and by definition of tensor spectral norm, we can bound the spectral norm of $\nabla^2 f(\mathbf{x})$ as

$$\begin{aligned}
\|\nabla^2 f(\mathbf{x})\| &\leq \sum_{k=2}^{d} k(k-1)\|\mathcal{A}_k \times_1 \mathbf{x} \times_2 \mathbf{x} \times_3 \mathbf{x} \cdots \times_{k-2} \mathbf{x}\| \\
&= \sum_{k=2}^{d} k(k-1) \max_{\|\mathbf{y}\|_2=1, \|\mathbf{z}\|_2=1} \langle \mathcal{A}_k \times_1 \mathbf{x} \times_2 \mathbf{x} \times_3 \mathbf{x} \cdots \times_{k-2} \mathbf{x}, \mathbf{y} \otimes \mathbf{z}\rangle \\
&= \sum_{k=2}^{d} k(k-1) \max_{\|\mathbf{y}\|_2=1, \|\mathbf{z}\|_2=1} \mathcal{A}_k \times_1 \mathbf{x} \times_2 \mathbf{x} \times_3 \mathbf{x} \cdots \times_{k-2} \mathbf{x} \times_{k-1} \mathbf{y} \times_k \mathbf{z} \\
&\leq \sum_{k=2}^{d} k(k-1)\|\mathcal{A}_k\|\|\mathbf{x}\|_2^{k-2}
\end{aligned} \tag{G.24}$$

where the second line follows from the definition of matrix spectral norm, the third line follows from the definition of kth-mode tensor-vector product (G.23), and the fourth line follows from the definition of a general Nth-order tensor, that is,

$$\|\mathcal{A}\| := \max_{\|\mathbf{x}_i\|_2=1 \forall i} \mathcal{A} \times_1 \mathbf{x}_1 \times_2 \mathbf{x}_2 \cdots \times_N \mathbf{x}_N.$$

Therefore, by (G.24), we have

$$\|\nabla^2 f(\mathbf{x})\| \leq \sum_{k=2}^{d} k(k-1)\|\mathcal{A}_k\|(1+\|\mathbf{x}\|_2^{d-2}) \tag{G.25}$$

Meanwhile, from the first line of (G.21), we have

$$\nabla^2 h(\mathbf{x}) \succeq (\alpha\|\mathbf{x}\|_2^{d-2}+\sigma)\mathbf{I}_n \tag{G.26}$$

Combing (G.25) and (G.26), we finally get that $(f(\mathbf{x}), h(\mathbf{x}))$ satisfies L_f-adaptive Lipschitz gradient condition for any

$$L_f \geq \sum_{k=2}^{d} k(k-1)\|\mathcal{A}_k\| \max\left\{\frac{1}{\sigma}, \frac{1}{\alpha}\right\}.$$

Showing Assumption 10.2.4. It directly follows from the coercivity assumption of f and Remark 10.2.2.

This completes the proof of showing that $(f(\mathbf{x}), h(\mathbf{x}))$ satisfies Assumptions 10.2.1- 10.2.4.

Showing $(f(\mathbf{x},\mathbf{y}), h(\mathbf{x},\mathbf{y}))$ satisfying Assumptions 10.2.1- 10.2.4. First, any (d_1, d_2)th-degree polynomial function $f(\mathbf{x},\mathbf{y})$ can be represented as

$$f(\mathbf{x},\mathbf{y}) = \sum_{i=0}^{d_1}\sum_{j=0}^{d_2}\langle \mathcal{A}_{i,j}, \mathbf{x}^{\otimes i}\otimes\mathbf{y}^{\otimes j}\rangle \tag{G.27}$$

where the coefficients of (i,j)th-order monomials are arranged as $\mathcal{A}_{i,j} \in \prod_{k=1}^{i}\mathbb{R}^n \times \prod_{k=1}^{j}\mathbb{R}^m$. For convenience, we denote $\mathbf{x}^{\otimes 0} = \mathbf{y}^{\otimes 0} = 1$ and $\mathcal{A}_{0,0} \in \mathbb{R}$. Further, due to super-symmetric tensors $\mathbf{x}^{\otimes i}$ and $\mathbf{y}^{\otimes j}$, we can always assume $\mathcal{A}_{i,j}$ for $i \geq 2$ or $j \geq 2$ as bi-super-symmetric tensors, i.e., those entries $\mathcal{A}_{i,j}(k_1, \cdots, k_i, k_{i+1}, \cdots, k_{i+j})$ have the same value despite the order of $(k_1, k_2, \cdots, k_j)$ and the order of $(k_{i+1}, k_{i+1}, \cdots, k_{i+j})$.

Due to the symmetric structures between $h(\mathbf{x},\mathbf{y})$ and $h(y)$ and by a similar argument as in the proof of showing $(f(\mathbf{x}), h(\mathbf{x}))$ satisfies Assumptions 10.2.1- 10.2.4, it suffices to show that

Part 1 $h(\mathbf{x},\mathbf{y})$ is bi-super-coercive and σ-strongly bi-convex;

Part 2 $(f(\mathbf{x},\mathbf{y}), h(\mathbf{x},\mathbf{y}))$ satisfies bi-adaptive Lipschitz gradient condition,

and the remaining parts of Assumptions 10.2.1- 10.2.4 can be directly obtained from the given assumptions on $f(\mathbf{x},\mathbf{y})$ and $h(\mathbf{x},\mathbf{y})$.

Showing Part 1. First of all, we observe that

$$\lim_{\|\mathbf{x}\|_2\to\infty} \frac{h(\mathbf{x},\mathbf{y})}{\|\mathbf{x}\|_2} \geq \lim_{\|\mathbf{x}\|_2\to\infty} \frac{\sigma}{2}\|\mathbf{x}\|_2 = \infty,$$
$$\lim_{\|\mathbf{y}\|_2\to\infty} \frac{h(\mathbf{x},\mathbf{y})}{\|\mathbf{y}\|_2} \geq \lim_{\|\mathbf{x}\|_2\to\infty} \frac{\sigma}{2}\|\mathbf{y}\|_2 = \infty,$$

which implies that $h(\mathbf{x},\mathbf{y})$ is bi-super-coercive. It remains to show that $h(\mathbf{x},\mathbf{y})$ is σ-strongly bi-convex. Towards that end, we compute the partial Hessians of $h(\mathbf{x},\mathbf{y})$. Similar to (G.20), we have

$$\nabla^2_{\mathbf{xx}} h(\mathbf{x},\mathbf{y}) = \left(\frac{\alpha}{d_2}\|\mathbf{y}\|_2^{d_2} + \frac{\sigma}{2}\|\mathbf{y}\|_2^2 + 1\right)\left(\alpha(d_1-2)\|\mathbf{x}\|_2^{d_1-2}\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\frac{\mathbf{x}^\top}{\|\mathbf{x}\|_2} + \left(\alpha\|\mathbf{x}\|_2^{d_1-2}+\sigma\right)\mathbf{I}_n\right) \tag{G.28}$$
$$\nabla^2_{\mathbf{yy}} h(\mathbf{x},\mathbf{y}) = \left(\frac{\alpha}{d_1}\|\mathbf{x}\|_2^{d_1} + \frac{\sigma}{2}\|\mathbf{x}\|_2^2 + 1\right)\left(\alpha(d_2-2)\|\mathbf{y}\|_2^{d_2-2}\frac{\mathbf{y}}{\|\mathbf{y}\|_2}\frac{\mathbf{y}^\top}{\|\mathbf{y}\|_2} + \left(\alpha\|\mathbf{y}\|_2^{d_2-2}+\sigma\right)\mathbf{I}_m\right) \tag{G.29}$$

This then implies that

$$\nabla^2_{\mathbf{xx}} h(\mathbf{x},\mathbf{y}) \succeq \sigma\mathbf{I}_n,$$
$$\nabla^2_{\mathbf{yy}} h(\mathbf{x},\mathbf{y}) \succeq \sigma\mathbf{I}_m.$$

Therefore, $h(\mathbf{x},\mathbf{y})$ is σ-strongly bi-convex. This completes the proof of Part 1.

Showing Part 2. We first bound the partial Hessians spectral norms of $f(\mathbf{x},\mathbf{y})$. Similar to (G.24), we bounded $\|\nabla^2_{\mathbf{xx}} f(\mathbf{x},\mathbf{y})\|$ as (using the bi-super-symmetry of $\mathcal{A}_{i,j}$)

$$\begin{aligned}\|\nabla^2_{\mathbf{xx}} f(\mathbf{x},\mathbf{y})\| &\leq \sum_{i=2}^{d_1}\sum_{j=0}^{d_2} i(i-1)\|\mathcal{A}_{i,j} \times_1 \mathbf{x} \times_2 \mathbf{x} \cdots \times_{i-2} \mathbf{x} \times_{i+1} \mathbf{y} \cdots \times_{i+j} \mathbf{y}\| \\ &\leq \sum_{i=2}^{d_1}\sum_{j=0}^{d_2} i(i-1)\|\mathcal{A}_{i,j}\|\|\mathbf{x}\|_2^{i-2}\|\mathbf{y}\|_2^j.\end{aligned} \tag{G.30}$$

Similarly, we bounded $\|\nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y})\|$ as

$$\begin{aligned}\|\nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y})\| &\leq \sum_{i=0}^{d_1}\sum_{j=2}^{d_2} j(j-1)\|\mathcal{A}_{i,j} \times_1 \mathbf{x} \times_2 \mathbf{x} \cdots \times_i \mathbf{x} \times_{i+1} \mathbf{y} \cdots \times_{i+j-2} \mathbf{y}\| \\ &\leq \sum_{i=0}^{d_1}\sum_{j=2}^{d_2} j(j-1)\|\mathcal{A}_{i,j}\|\|\mathbf{x}\|_2^{i}\|\mathbf{y}\|_2^{j-2}.\end{aligned} \tag{G.31}$$

Now, similar to (G.25), we obtain from (G.30) and (G.31) that

$$\|\nabla^2_{\mathbf{xx}} f(\mathbf{x},\mathbf{y})\| \leq (1+\|\mathbf{x}\|_2^{d_1-2})(1+\|\mathbf{y}\|_2^{d_2})\sum_{i=2}^{d_1}\sum_{j=0}^{d_2} i(i-1)\|\mathcal{A}_{i,j}\|, \tag{G.32}$$

$$\|\nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y})\| \leq (1+\|\mathbf{x}\|_2^{d_1})(1+\|\mathbf{y}\|_2^{d_2-2})\sum_{i=0}^{d_1}\sum_{j=2}^{d_2} j(j-1)\|\mathcal{A}_{i,j}\| \tag{G.33}$$

Further, using (G.28) and (G.29), we can further bound

$$\nabla^2_{\mathbf{xx}} h(\mathbf{x},\mathbf{y}) \succeq \left(\alpha\|\mathbf{x}\|_2^{d_1-2}+\sigma\right)\left(\frac{\alpha}{d_2}\|\mathbf{y}\|_2^{d_2}+1\right)\mathbf{I}_n, \tag{G.34}$$

$$\nabla^2_{\mathbf{yy}} h(\mathbf{x},\mathbf{y}) \succeq \left(\frac{\alpha}{d_1}\|\mathbf{x}\|_2^{d_1}+1\right)\left(\alpha\|\mathbf{y}\|_2^{d_2-2}+\sigma\right)\mathbf{I}_m \tag{G.35}$$

Then, combining (G.32)- (G.35), we have $(f(\mathbf{x},\mathbf{y}), h(\mathbf{x},\mathbf{y}))$ satisfies (L_1, L_2)-bi-adaptive Lipschitz gradient condition for any

$$\begin{aligned} L_1 &\geq \frac{1}{\sum_{i=2}^{d_1}\sum_{j=0}^{d_2} i(i-1)\|\mathcal{A}_{i,j}\|}\max\left\{1,\frac{1}{\sigma},\frac{1}{\alpha},\frac{d_2}{\alpha}\right\} \\ &= \frac{1}{\sum_{i=2}^{d_1}\sum_{j=0}^{d_2} i(i-1)\|\mathcal{A}_{i,j}\|}\max\left\{1,\frac{1}{\sigma},\frac{d_2}{\alpha}\right\}, \end{aligned}$$

and

$$\begin{aligned} L_2 &\geq \frac{1}{\sum_{i=0}^{d_1}\sum_{j=2}^{d_2} j(j-1)\|\mathcal{A}_{i,j}\|}\max\left\{1,\frac{1}{\alpha},\frac{1}{\sigma},\frac{d_1}{\alpha}\right\} \\ &= \frac{1}{\sum_{i=0}^{d_1}\sum_{j=2}^{d_2} j(j-1)\|\mathcal{A}_{i,j}\|}\max\left\{1,\frac{1}{\sigma},\frac{d_1}{\alpha}\right\}. \end{aligned}$$

This completes the argument of showing Part 2,

Combining all the above, we finish the proof of showing that $(f(\mathbf{x},\mathbf{y}), h(\mathbf{x},\mathbf{y}))$ satisfies Assumptions 10.2.1-10.2.4. □

G.3.2 Application to Any Objective Functions with a Polynomial-order Hessian Spectral Norm

Lemma G.3.2 (Lemma 10.3.2)**.** *Suppose* $\|\nabla^2 f(\mathbf{x})\| \leq C_1 + C_2\|\mathbf{x}\|_2^{d-2}$ *(or* $\|\nabla^2_{\mathbf{xx}} f(\mathbf{x},\mathbf{y})\| \leq (C_1+C_2\|\mathbf{x}\|_2^{d_1-2})(C_3+C_4\|\mathbf{y}\|_2^{d_2})$ *and* $\|\nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y})\| \leq (C_5+C_6\|\mathbf{x}\|_2^{d_1})(C_7+C_8\|\mathbf{y}\|_2^{d_2-2})$*) in the whole domain with* $d, d_1, d_2 \geq 2$ *for some positive constants* C_1 *to* C_8*. Set the Bregman (or bi-Bregman) distance kernel* h *according to* (10.22) *for any* $\alpha, \sigma > 0$*. Then* (f, h) *(or* $(f(\mathbf{x},\mathbf{y}), h(\mathbf{x},\mathbf{y}))$*) satisfies the* L_f*-adaptive (or* (L_1, L_2)*-bi-adaptive) Lipschitz gradient condition for any* $L_f \geq \max\{\frac{C_1}{\sigma}, \frac{C_2}{\alpha}\}$ *and any* $L_1 \geq \max\{\frac{C_1}{\sigma}, \frac{C_2}{\alpha}, C_3, \frac{C_4 d_2}{\alpha}\}$ *and any* $L_2 \geq \max\{C_5, \frac{C_6 d_1}{\alpha}, \frac{C_7}{\sigma}, \frac{C_8}{\sigma}\}$.

Proof. We will show the adaptive Lipschitz condition and bi-adaptive Lipschitz condition, respectively.

Showing the adaptive Lipschitz condition. By definition of adaptive Lipschitz condition, it suffices to show that there is a $L_f > 0$ such that

$$L_f \nabla^2 h(\mathbf{x}) \pm \nabla^2 f(\mathbf{x}) \succeq 0$$

in the whole domain. In one way, by assumption of $f(\mathbf{x})$, we have

$$\|\nabla^2 f(\mathbf{x})\| \leq C_1 + C_2\|\mathbf{x}\|_2^{d-2}$$

in the whole domain with $d \geq 2$ for some positive constants C_1, C_2. In another way, by direct computations, $h(\mathbf{x})$ in (10.22) satisfies that

$$\nabla^2 h(\mathbf{x}) \succeq (\alpha\|\mathbf{x}\|_2^{d-2} + \sigma)\mathbf{I}_n \text{ for any } d \geq 2,$$

in the whole domain. Therefore, it is clear to see that

$$L_f \nabla^2 h(\mathbf{x}) \pm \nabla^2 f(\mathbf{x}) \succeq 0$$

in the whole domain for any $L_f \geq \max\{\frac{C_1}{\sigma}, \frac{C_2}{\alpha}\}$.

Showing the bi-adaptive Lipschitz condition. By definition of adaptive Lipschitz condition, it suffices to show that there are $L_1, L_2 > 0$ such that

$$\begin{aligned} L_1 \nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y}) \pm \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) \succeq 0, \\ L_2 \nabla^2_{\mathbf{yy}} h(\mathbf{x}, \mathbf{y}) \pm \nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \succeq 0. \end{aligned}$$

in the whole domain. In one way, by assumption of $f(\mathbf{x}, \mathbf{y})$, we have

$$\begin{aligned} \|\nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y})\| \leq \left(C_1 + C_2\|\mathbf{x}\|_2^{d_1-2}\right)\left(C_3 + C_4\|\mathbf{y}\|_2^{d_2}\right) \\ \|\nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y})\| \leq \left(C_5 + C_6\|\mathbf{x}\|_2^{d_1}\right)\left(C_7 + C_8\|\mathbf{y}\|_2^{d_2-2}\right) \end{aligned}$$

in the whole domain with $d_1, d_2 \geq 2$ for some positive constants C_1 to C_8. In another way, by direct computations, $h(\mathbf{x}, \mathbf{y})$ in (10.22) satisfies that for any $d_1, d_2 \geq 2$,

$$\begin{aligned} \nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y}) \succeq \left(\alpha\|\mathbf{x}\|_2^{d_1-2} + \sigma\right)\left(\frac{\alpha}{d_2}\|\mathbf{y}\|_2^{d_2} + \frac{\sigma}{2}\|\mathbf{y}\|_2^2 + 1\right)\mathbf{I}_n \succeq \left(\alpha\|\mathbf{x}\|_2^{d_1-2} + \sigma\right)\left(\frac{\alpha}{d_2}\|\mathbf{y}\|_2^{d_2} + 1\right)\mathbf{I}_n \\ \nabla^2_{\mathbf{yy}} h(\mathbf{x}, \mathbf{y}) \succeq \left(\frac{\alpha}{d_1}\|\mathbf{x}\|_2^{d_1} + \frac{\sigma}{2}\|\mathbf{x}\|_2^2 + 1\right)\left(\alpha\|\mathbf{y}\|_2^{d_2-2} + \sigma\right)\mathbf{I}_n \succeq \left(\frac{\alpha}{d_1}\|\mathbf{x}\|_2^{d_1} + 1\right)\left(\alpha\|\mathbf{y}\|_2^{d_2-2} + \sigma\right)\mathbf{I}_n \end{aligned}$$

Therefore, it is clear to see that

$$\begin{aligned} L_1 \nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y}) \pm \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) \succeq 0, \\ L_2 \nabla^2_{\mathbf{yy}} h(\mathbf{x}, \mathbf{y}) \pm \nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \succeq 0. \end{aligned}$$

for any L_1 and L_2 satifying

$$\begin{aligned} L_1 \geq \max\left\{\frac{C_1}{\sigma}, \frac{C_2}{\alpha}, C_3, \frac{C_4 d_2}{\alpha}\right\}, \\ L_2 \geq \max\left\{C_5, \frac{C_6 d_1}{\alpha}, \frac{C_7}{\sigma}, \frac{C_8}{\sigma}\right\}. \end{aligned}$$

□

G.4 Analysis of Algorithms 3-6

G.4.1 Convergence Analysis of Algorithm 3

For completeness of the proof, we still collect the convergence analysis of Algorithm 3 here, despite providing it in the main context.

Algorithm 10 Bregman Gradient Descent

1: **Input:** A Bregman kernel h with $L_f\nabla^2 h(\mathbf{x}) \pm \nabla^2 f(\mathbf{x}) \succeq 0$ in the whole domain; Set $\eta \in (0, \frac{1}{L_f})$.

2: **Initialization:** $\mathbf{x}^0$

3: **Recursion:** Iteratively generate a sequence $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ via

$$\mathbf{x}^\ell = g(\mathbf{x}^{\ell-1}) := \arg\min_{\mathbf{x}} \langle \nabla f(\mathbf{x}^{\ell-1}), \mathbf{x} - \mathbf{x}^\ell\rangle + \frac{1}{\eta} D_h(\mathbf{x}, \mathbf{x}^{\ell-1}) \tag{10.18}$$

G.4.1.1 First-order Convergence of Algorithm 3

Theorem G.4.1. *Under Assumptions 10.2.1– 10.2.4, Algorithm 3 with arbitrary initialization converges to a critical point of f in* (10.1)*.*

Proof. First, it is clear that Algorithm 3 is well-defined in view of Lemma 10.2.1. Then in view of Theorem 10.4.1 and the assumption that f is KL function, it is sufficient to prove that $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ is a gradient-like descent sequence for f (see Definition 10.4.1), i.e., to show:

(C1) Sufficient decrease property: $f(\mathbf{x}^\ell) - f(\mathbf{x}^{\ell+1}) \geq \rho_1 \left\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\right\|_2^2,\ \forall\, \ell \in \mathbb{N}$ for some $\rho_1 > 0$;

(C2) Bounded gradient property: $\left\|\nabla f(\mathbf{x}^{\ell+1})\right\|_2 \leq \rho_2 \left\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\right\|_2,\ \forall\, \ell \in \mathbb{N}$ for some $\rho_2 > 0$.

Condition (C1) follows from (10.10) in Lemma 10.2.1. Condition (C2) holds because by the optimality condition

$$\nabla f(\mathbf{x}^\ell) + (\nabla h(\mathbf{x}^{\ell+1}) - \nabla h(\mathbf{x}^\ell))/\eta = \mathbf{0}, \tag{G.36}$$

we have

$$\|\nabla f(\mathbf{x}^\ell)\|_2 = \frac{1}{\eta}\|\nabla h(\mathbf{x}^{\ell+1}) - \nabla h(\mathbf{x}^\ell)\|_2 \leq \frac{\rho_h(\mathcal{B})}{\eta}\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\|_2,$$

where the inequality follows from Assumption 10.2.4, $h \in \mathcal{C}^2$, and the fact any function in $\mathcal{C}^2$ admits a locally Lipschitz gradient on any bounded set (see Footnote 40). Therefore, by continuing this argument, we claim that f has a locally $\rho_f(\mathcal{B})$-Lipschitz gradient on $\mathcal{B}$, and we have $\|\nabla f(\mathbf{x}^{\ell+1})\|_2 \leq \left(\frac{\rho_h(\mathcal{B})}{\eta} + \rho_f(\mathcal{B})\right)\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\|_2$. □

G.4.1.2 Second-order Convergence of Algorithm 3

Theorem G.4.2. *Under Assumptions 10.2.1– 10.2.4, Algorithm 3 with random initialization almost surely converges to a second-order stationary point of f in* (10.1)*.*

Proof. To show the second-order convergence from the first-order convergence, it suffices to show that Algorithm 3 avoids strict saddles. We define (10.18) as $\mathbf{x}^\ell = g(\mathbf{x}^{\ell-1})$ and compute the Jacobian Dg. By the definition of g, we get $Dg(\mathbf{x}^\ell) = \partial \mathbf{x}^{\ell+1}/\partial \mathbf{x}^\ell$. Then we apply the implicit function theorem to the optimality condition (G.36) and in view of the nonsingularity of $\nabla^2 h$, we obtain that Dg is continuous and given by

$$Dg(\mathbf{x}^\ell) = \left[\nabla^2 h(\mathbf{x}^{\ell+1})\right]^{-1} (\nabla^2 h(\mathbf{x}^\ell) - \eta \nabla^2 f(\mathbf{x}^\ell)).$$

Since the above analysis holds for all $\mathbf{x}^\ell \in \mathbb{R}^n$, this further implies that $Dg(\mathbf{x})$ is continuous and given by

$$Dg(\mathbf{x}) = \left[\nabla^2 h(g(\mathbf{x}))\right]^{-1} \left(\nabla^2 h(\mathbf{x}) - \eta \nabla^2 f(\mathbf{x})\right). \tag{G.37}$$

To show the avoidance of strict saddles, by Theorem 10.4.2, it suffices to show the following conditions:

Showing g is a $\mathcal{C}^1$ mapping. This follows from the continuity of Dg in (G.37).

Showing $\det(Dg) \neq 0$ in the whole domain. By the positive definiteness of $\nabla^2 h$ and $\nabla^2 h \pm \eta \nabla^2 f$,

$$\det(Dg(\mathbf{x})) = \det([\nabla^2 h(g(\mathbf{x}))]^{-1}) \det(\nabla^2 h(\mathbf{x}) - \eta \nabla^2 f(\mathbf{x})) > 0.$$

Showing any strict saddle of f lies in $\mathcal{A}_g$. First for any strict saddle $\mathbf{x}^\star$, we have $\mathbf{x}^{\ell+1} = \mathbf{x}^\ell = \mathbf{x}^\star$ satisfies the optimality condition (G.36), so $\mathbf{x}^\star$ is a fixed point, i.e., $g(\mathbf{x}^\star) = \mathbf{x}^\star$. Plugging $g(\mathbf{x}^\star) = \mathbf{x}^\star$ into (G.37):

$$\begin{aligned} Dg(\mathbf{x}^\star) =& [\nabla^2 h(\mathbf{x}^\star)]^{-1} (\nabla^2 h(\mathbf{x}^\star) - \eta \nabla^2 f(\mathbf{x}^\star)) \\ \sim& [\nabla^2 h(\mathbf{x}^\star)]^{-\frac{1}{2}} (\nabla^2 h(\mathbf{x}^\star) - \eta \nabla^2 f(\mathbf{x}^\star)) [\nabla^2 h(\mathbf{x}^\star)]^{-\frac{1}{2}} \\ =& \mathbf{I} - \eta [\nabla^2 h(\mathbf{x}^\star)]^{-\frac{1}{2}} \nabla^2 f(\mathbf{x}^\star) [\nabla^2 h(\mathbf{x}^\star)]^{-\frac{1}{2}} := \mathbf{I} - \eta \mathbf{\Phi} \end{aligned}$$

with "$\sim$" denotes the matrix similarity. Therefore, $Dg(\mathbf{x}^\star)$ has an eigenvalue strictly greater than 1 since $\mathbf{\Phi}$ has a negative eigenvalue. This is because $\mathbf{\Phi}$ is congruent to $\nabla^2 f(\mathbf{x}^\star)$, which has a negative eigenvalue. □

G.4.2 Convergence Analysis of Algorithm 4

Algorithm 11 Bregman Alternating Gradient Descent

1: **Input:** A bi-Bregman kernel $h(\mathbf{x},\mathbf{y})$ with both $L_1\nabla^2_{\mathbf{xx}}h(\mathbf{x},\mathbf{y}) \pm \nabla^2_{\mathbf{xx}}f(\mathbf{x},\mathbf{y}) \succeq 0$ and $L_2\nabla^2_{\mathbf{yy}}h(\mathbf{x},\mathbf{y}) \pm \nabla^2_{\mathbf{yy}}f(\mathbf{x},\mathbf{y}) \succeq 0$ in the entire domain; Set $\eta \in (0, \min(\frac{1}{L_1}, \frac{1}{L_2}))$.

2: **Initialization:** $(\mathbf{x}^0, \mathbf{y}^0)$

3: **Recursion:** Iteratively generate a sequence $\{\mathbf{x}^\ell, \mathbf{y}^\ell\}_{\ell\in\mathbb{N}}$ via

$$\begin{aligned}\mathbf{x}^\ell &= \arg\min_{\mathbf{x}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}), \mathbf{x} - \mathbf{x}^{\ell-1}\rangle + \frac{1}{\eta} D_h^1(\mathbf{x}, \mathbf{x}^{\ell-1}; \mathbf{y}^{\ell-1}),\\ \mathbf{y}^\ell &= \arg\min_{\mathbf{y}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}^{\ell}, \mathbf{y}^{\ell-1}), \mathbf{y} - \mathbf{y}^{\ell-1}\rangle + \frac{1}{\eta} D_h^2(\mathbf{y}, \mathbf{y}^{\ell-1}; \mathbf{x}^{\ell})\end{aligned} \tag{G.38}$$

G.4.2.1 First-order Convergence of Algorithm 4

Theorem G.4.3. *Under Assumptions 10.2.1– 10.2.4, Algorithm 4 with arbitrary initialization converges to a critical point of f in* (10.12)*.*

Proof. First of all, in view of Lemma 10.2.2, we immediately conclude that Algorithm 4 is well-defined:

Proposition G.4.1. *Under Assumptions 10.2.1– 10.2.4, Algorithm 4 is well-defined.*

Now, by Theorem 10.4.1 and the assumption that f is KL function, it is sufficient to prove that $\{(\mathbf{x}^\ell, \mathbf{y}^\ell)\}_{\ell\in\mathbb{N}}$ is a gradient-like descent sequence for f (see Definition 10.4.1), i.e., showing:

(C1) Sufficient decrease property;

(C2) Bounded gradient property.

Condition (C1) directly follows from Lemma 10.2.2.

To show Condition (C2), we start with the optimality condition (G.39) for the first-block of Algorithm 4

$$\nabla_{\mathbf{x}} h(\mathbf{x}^+, \mathbf{y}) = \nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{y}) - \eta \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \tag{G.39}$$

which implies

$$\begin{aligned}\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|_2 &= \frac{1}{\eta}\|\nabla_{\mathbf{x}} h(\mathbf{x}^+,\mathbf{y}) - \nabla_{\mathbf{x}} h(\mathbf{x},\mathbf{y})\|_2\\ &\le \frac{\rho_h(\mathcal{B})}{\eta}\|\mathbf{x}^+ - \mathbf{x}\|_2\\ &\le \frac{\rho_h(\mathcal{B})}{\eta}\|(\mathbf{x}^+,\mathbf{y}^+) - (\mathbf{x},\mathbf{y})\|_2\end{aligned}$$

where the second line follows from Assumption 10.2.4, $h \in \mathcal{C}^2$, and Footnote 40.

Using a similar analysis for the optimality condition for the second-block of Algorithm 4

$$\nabla_{\mathbf{y}} h(\mathbf{x}^+, \mathbf{y}^+) = \nabla_{\mathbf{y}} h(\mathbf{x}^+, \mathbf{y}) - \eta \nabla_{\mathbf{y}} f(\mathbf{x}^+, \mathbf{y}), \tag{G.40}$$

we can get

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}^+, \mathbf{y})\|_2 \leq \frac{\rho_h(\mathcal{B})}{\eta} \|\mathbf{y}^+ - \mathbf{y}\|_2$$

implying

$$\begin{aligned}
\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|_2 &\leq \|\nabla_{\mathbf{y}} f(\mathbf{x}^+, \mathbf{y})\|_2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}^+, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|_2 \\
&\leq \frac{\rho_h(\mathcal{B})}{\eta} \|\mathbf{y}^+ - \mathbf{y}\|_2 + \rho_f(\mathcal{B}) \|\mathbf{x}^+ - \mathbf{x}\|_2 \\
&\leq \left(\frac{\rho_h(\mathcal{B})}{\eta} + \rho_f(\mathcal{B}) \right) \|(\mathbf{x}^+, \mathbf{y}^+) - (\mathbf{x}, \mathbf{y})\|_2
\end{aligned}$$

where the second line follows from Assumption 10.2.4, $f \in \mathcal{C}^2$, and Footnote 40.

Combing the above two, we get an equivalent version of the bounded gradient property

$$\begin{aligned}
\|\nabla f(\mathbf{x}, \mathbf{y})\|_2 \leq & \|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}\|_2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|_2 \\
&\leq \left(\frac{2\rho_h(\mathcal{B})}{\eta} + \rho_f(\mathcal{B}) \right) \|(\mathbf{x}^+, \mathbf{y}^+) - (\mathbf{x}, \mathbf{y})\|_2
\end{aligned}$$

Therefore,

$$\|\nabla f(\mathbf{x}^+, \mathbf{y}^+)\|_2 \leq \left(\frac{2\rho_h(\mathcal{B})}{\eta} + 2\rho_f(\mathcal{B}) \right) \|(\mathbf{x}^+, \mathbf{y}^+) - (\mathbf{x}, \mathbf{y})\|_2.$$

□

G.4.2.2 Second-order Convergence of Algorithm 4

Theorem G.4.4. *Under Assumptions 10.2.1– 10.2.4, Algorithm 4 with random initialization almost surely converges to a second-order stationary point of f in Equation* (10.12)*.*

Proof. Following from (G.38), we denote

$$\begin{aligned}
(\mathbf{x}^+, \mathbf{y}) &= g_1(\mathbf{x}, \mathbf{y}) \\
(\mathbf{x}, \mathbf{y}^+) &= g_2(\mathbf{x}, \mathbf{y})
\end{aligned} \tag{G.41}$$

The mappings g_1, g_2 are well-defined in the whole domain $\mathbb{R}^n \times \mathbb{R}^m$ in view of strong convexity and coercivity of the objective function in (G.38). Then Algorithm 4 can be viewed as iteratively performing the following composite mapping for $\ell = 1, 2, \ldots$

$$(\mathbf{x}^\ell, \mathbf{y}^\ell) = g(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}) \tag{G.42}$$

where the mapping g is defined as the composite mapping of g_1, g_2:

$$g := g_2 \circ g_1 \tag{G.43}$$

To compute the Jacobian matrix Dg, we first compute Dg_1 and Dg_2. Then we will use the chain rule to get Dg.

Now we compute $Dg_1(\mathbf{x}, \mathbf{y})$ using $(\mathbf{x}^+, \mathbf{y}) = g_1(\mathbf{x}, \mathbf{y})$. First of all,

$$Dg_1(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \frac{\partial \mathbf{x}^+}{\partial \mathbf{x}^\top} & \frac{\partial \mathbf{x}^+}{\partial \mathbf{y}^\top} \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix}$$

Then by the first-order optimality condition (G.39) of Algorithm 4:

$$\nabla_{\mathbf{x}} h(\mathbf{x}^+, \mathbf{y}) = \nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{y}) - \eta \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$$

Now apply the Implicit function theorem to (G.39)

$$\begin{aligned} \nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y}) \frac{\partial \mathbf{x}^+}{\partial \mathbf{x}^\top} &= \nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y}) - \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) \\ \nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y}) \frac{\partial \mathbf{x}^+}{\partial \mathbf{y}^\top} &= \nabla^2_{\mathbf{xy}} h(\mathbf{x}, \mathbf{y}) - \nabla^2_{\mathbf{xy}} h(\mathbf{x}^+, \mathbf{y}) - \eta \nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y}) \end{aligned}$$

which implies (since $\nabla^2 h_1$ is positive definite in the whole domain) the following Jacobians are continuous and given by

$$\begin{aligned} \frac{\partial \mathbf{x}^+}{\partial \mathbf{x}^\top} &= \nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y})^{-1} \left(\nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y}) - \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) \right) \\ \frac{\partial \mathbf{x}^+}{\partial \mathbf{y}^\top} &= \nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y})^{-1} (\nabla^2_{\mathbf{xy}} h(\mathbf{x}, \mathbf{y}) - \nabla^2_{\mathbf{xy}} h(\mathbf{x}^+, \mathbf{y}) - \eta \nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y})) \end{aligned}$$

Therefore, Dg_1 is continuous and given by

$$\begin{aligned} Dg_1(\mathbf{x}, \mathbf{y}) &= \begin{bmatrix} \frac{\partial \mathbf{x}^+}{\partial \mathbf{x}} & \frac{\partial \mathbf{x}^+}{\partial \mathbf{y}^\top} \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix} \\ &= \begin{bmatrix} \nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix} \\ &\quad \begin{bmatrix} \nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y}) - \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) & \nabla^2_{\mathbf{xy}} h(\mathbf{x}, \mathbf{y}) - \nabla^2_{\mathbf{xy}} h(\mathbf{x}^+, \mathbf{y}) - \eta \nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y}) \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix} \end{aligned} \tag{G.44}$$

Similarly, we have Dg_2 is continuous and given by

$$
\begin{aligned}
Dg_2(\mathbf{x},\mathbf{y}) &= \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \frac{\partial \mathbf{y}^+}{\partial \mathbf{x}^\top} & \frac{\partial \mathbf{y}^+}{\partial \mathbf{y}^\top} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \nabla^2_{\mathbf{yy}} h(\mathbf{x},\mathbf{y}^+)^{-1} \end{bmatrix} \\
&\quad \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \nabla^2_{\mathbf{yx}} h(\mathbf{x},\mathbf{y}) - \nabla^2_{\mathbf{yx}} h(\mathbf{x},\mathbf{y}^+) - \eta \nabla^2_{\mathbf{yx}} f(\mathbf{x},\mathbf{y}) & \nabla^2_{\mathbf{yy}} h(\mathbf{x},\mathbf{y}) - \eta \nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y}) \end{bmatrix}
\end{aligned} \tag{G.45}
$$

Finally, using the chain rule, we get Dg is continuous (since continuity is closed under composite and product operations) and given by

$$
Dg(\mathbf{x},\mathbf{y}) = Dg_2(g_1(\mathbf{x},\mathbf{y}))Dg_1(\mathbf{x},\mathbf{y}). \tag{G.46}
$$

By Theorem 10.4.2, to show that the mapping g can almost surely avoid the strict saddles, it suffices to show the following conditions:

1. g is a $\mathcal{C}^1$ mapping;
2. $\det(Dg) \neq 0$ in the whole domain;
3. Any strict saddle of f is an unstable fixed point of g.

Showing g is $\mathcal{C}^1$ mapping. This follows from the continuity of Dg in (G.46).

Showing $\det(Dg) \neq 0$ in the whole domain. To show $\det(Dg) \neq 0$ in the whole domain, by using the chain rule

$$
Dg = Dg_2 Dg_1
$$

and noting that each Dg_k is a squared matrix, it suffices to show both Dg_1 and Dg_2 are nonsingular in the whole domain. Since Dg_1 is a block upper-triangular matrix (see (G.44)), it suffices to show both of its diagonal block matrices are nonsingular. The first diagonal block is $\nabla^2_{\mathbf{xx}} h(\mathbf{x}^+,\mathbf{y})^{-1}(\nabla^2_{\mathbf{xx}} h(\mathbf{x},\mathbf{y}) - \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x},\mathbf{y}))$, which is nonsingular in the whole domain because of h is strongly bi-convex (implying $\nabla^2_{\mathbf{xx}} h$ is positive definite in the whole domain and hence nonsingular in the whole domain) and (f,h) satisfies (L_1,L_2)-bi-adaptive Lipschitz gradient condition (implying $\nabla^2_{\mathbf{xx}} h(\mathbf{x},\mathbf{y}) - \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x},\mathbf{y})$ is positive definite in the whole domain for any $\eta < \frac{1}{L_1}$ and hence nonsingular in the whole domain). Therefore, we obtain that Dg_1 is nonsingular in the whole domain. Using a similar analysis and in view of (G.45), we can show that $Dg_2(\mathbf{x},\mathbf{y})$ is nonsingular in the whole domain. This shows that $\det(Dg) \neq 0$ in the whole domain.

Showing any strict saddle of f lies in $\mathcal{A}_g$. By definition, we want to show each strict saddle $(\mathbf{x}^\star,\mathbf{y}^\star)$ of f satisfies

1. $g(\mathbf{x}^\star,\mathbf{y}^\star) = (\mathbf{x}^\star,\mathbf{y}^\star)$

2. $\max_i |\lambda_i(Dg(\mathbf{x}^\star, \mathbf{y}^\star))| > 1$

The first part comes from that $(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^+, \mathbf{y}^+) = (\mathbf{x}^\star, \mathbf{y}^\star)$ satisfies the first-order optimality condition of (G.38):

$$\begin{aligned}\nabla_{\mathbf{x}} h(\mathbf{x}^\star, \mathbf{y}^\star) &= \nabla_{\mathbf{x}} h(\mathbf{x}^\star, \mathbf{y}^\star) - \eta \nabla_{\mathbf{x}} f(\mathbf{x}^\star, \mathbf{y}^\star) \\ \nabla_{\mathbf{y}} h(\mathbf{x}^\star, \mathbf{y}^\star) &= \nabla_{\mathbf{y}} h(\mathbf{x}^\star, \mathbf{y}^\star) - \eta \nabla_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star)\end{aligned}$$

which definitely hold since any strict saddle $(\mathbf{x}^\star, \mathbf{y}^\star)$ is also a critical point of f:

$$(\nabla_{\mathbf{x}} f(\mathbf{x}^\star), \nabla_{\mathbf{y}} f(\mathbf{x}^\star, \mathbf{y}^\star)) = (\mathbf{0}, \mathbf{0})$$

Further, $(\mathbf{x}^\star, \mathbf{y}^\star)$ is the unique point satisfying the above first-order optimality condition by Proposition G.4.1.

Now we show the second part, that is, the maximum eigenvalue of magnitude for $Dg(\mathbf{x}^\star, \mathbf{y}^\star)$ is greater than 1. To simplify notations, we make the following notations;

$$\begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} := \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\mathbf{x}^\star, \mathbf{y}^\star) & \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star, \mathbf{y}^\star) \\ \nabla^2_{\mathbf{yx}} f(\mathbf{x}^\star, \mathbf{y}^\star) & \nabla^2_{\mathbf{yy}} f(\mathbf{x}^\star, \mathbf{y}^\star) \end{bmatrix}$$

and

$$\begin{aligned}\mathbf{H}_1 &:= \nabla^2_{\mathbf{xx}} h(\mathbf{x}^\star, \mathbf{y}^\star) \\ \mathbf{H}_2 &:= \nabla^2_{\mathbf{yy}} h(\mathbf{x}^\star, \mathbf{y}^\star).\end{aligned}$$

Now we are ready to compute $Dg(\mathbf{x}^\star, \mathbf{y}^\star)$ by plugging

$$(\mathbf{x}^+, \mathbf{y}^+) = (\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\star, \mathbf{y}^\star)$$

to (G.46) and using the above notations:

$$\begin{aligned}Dg(\mathbf{x}^\star, \mathbf{y}^\star) =& \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ -\eta \mathbf{H}_2^{-1} \mathbf{F}_{21} & \mathbf{I}_m - \eta \mathbf{H}_2^{-1} \mathbf{F}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n - \eta \mathbf{H}_1^{-1} \mathbf{F}_{11} & -\eta \mathbf{H}_1^{-1} \mathbf{F}_{12} \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix} \\ =& \left(\mathbf{I} - \eta \begin{bmatrix} \mathbf{0} & \\ & \nabla^2 h_2(\mathbf{y}^\star)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} \right) \left(\mathbf{I} - \eta \begin{bmatrix} \mathbf{H}_1^{-1} & \\ & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} \right) \\ =& \mathbf{I} - \eta \begin{bmatrix} \mathbf{H}_1^{-1} & \\ & \mathbf{H}_2^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} + \eta^2 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{H}_2^{-1} \mathbf{F}_{21} \mathbf{H}_1^{-1} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} \\ =& \mathbf{I} - \begin{bmatrix} \eta \mathbf{H}_1^{-1} & \\ -\eta^2 \mathbf{H}_2^{-1} \mathbf{F}_{21} \mathbf{H}_1^{-1} & \eta \mathbf{H}_2^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} \\ =& \mathbf{I} - \begin{bmatrix} \frac{1}{\eta} \mathbf{H}_1 & \\ \mathbf{F}_{21} & \frac{1}{\eta} \mathbf{H}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} \\ =& \begin{bmatrix} \frac{1}{\eta} \mathbf{H}_1 & \\ \mathbf{F}_{21} & \frac{1}{\eta} \mathbf{H}_2 \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\eta} \mathbf{H}_1 - \mathbf{F}_{11} & -\mathbf{F}_{12} \\ & \frac{1}{\eta} \mathbf{H}_2 - \mathbf{F}_{22} \end{bmatrix}\end{aligned}$$

This implies

$$Dg(\mathbf{x}^\star, \mathbf{y}^\star)^{-1} = \begin{bmatrix} \frac{1}{\eta} \mathbf{H}_1 - \mathbf{F}_{11} & -\mathbf{F}_{12} \\ & \frac{1}{\eta} \mathbf{H}_2 - \mathbf{F}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\eta} \mathbf{H}_1 & \\ \mathbf{F}_{21} & \frac{1}{\eta} \mathbf{H}_2 \end{bmatrix}$$

Note that to show that $Dg(\mathbf{x}^\star, \mathbf{y}^\star)$ has an eigenvalue with magnitude larger than 1, it suffices to show that the equation

$$\det(Dg(\mathbf{x}^\star, \mathbf{y}^\star)^{-1} - \mu \mathbf{I}) = 0$$

has a solution $\mu \in (0, 1)$.

Towards that end, note that $\det(Dg(\mathbf{x}^\star, \mathbf{y}^\star)^{-1} - \mu \mathbf{I}) = 0$ is equivalent to

$$\begin{aligned}
&\det\left(\begin{bmatrix} \frac{1}{\eta}\mathbf{H}_1 - \mathbf{F}_{11} & -\mathbf{F}_{12} \\ & \frac{1}{\eta}\mathbf{H}_2 - \mathbf{F}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\eta}\mathbf{H}_1 & \\ \mathbf{F}_{21} & \frac{1}{\eta}\mathbf{H}_2 \end{bmatrix} - \mu \mathbf{I}\right) = 0 \\
\iff &\det\left(\begin{bmatrix} \frac{1}{\eta}\mathbf{H}_1 & \\ \mathbf{F}_{21} & \frac{1}{\eta}\mathbf{H}_2 \end{bmatrix} - \mu \begin{bmatrix} \frac{1}{\eta}\mathbf{H}_1 - \mathbf{F}_{11} & -\mathbf{F}_{12} \\ & \frac{1}{\eta}\mathbf{H}_2 - \mathbf{F}_{22} \end{bmatrix}\right) = 0 \\
\iff &\det\left(\begin{bmatrix} \frac{1}{\eta}(1-\mu)\mathbf{H}_1 + \mu\mathbf{F}_{11} & \mu\mathbf{F}_{12} \\ \mathbf{F}_{21} & \frac{1}{\eta}(1-\mu)\mathbf{H}_2 + \mu\mathbf{F}_{22} \end{bmatrix}\right) = 0 \\
\iff &\det\left(\begin{bmatrix} \sqrt{\mu}\mathbf{I}_n & \\ & \mathbf{I}_m \end{bmatrix}\right) \det\left(\begin{bmatrix} \frac{1}{\eta}(1-\mu)\mathbf{H}_1 + \mu\mathbf{F}_{11} & \sqrt{\mu}\mathbf{F}_{12} \\ \sqrt{\mu}\mathbf{F}_{21} & \frac{1}{\eta}(1-\mu)\mathbf{H}_2 + \mu\mathbf{F}_{22} \end{bmatrix}\right) \det\left(\begin{bmatrix} \frac{1}{\sqrt{\mu}}\mathbf{I}_n & \\ & \mathbf{I}_m \end{bmatrix}\right) = 0 \\
\iff &\det\left(\begin{bmatrix} \frac{1}{\eta}(1-\mu)\mathbf{H}_1 + \mu\mathbf{F}_{11} & \sqrt{\mu}\mathbf{F}_{12} \\ \sqrt{\mu}\mathbf{F}_{21} & \frac{1}{\eta}(1-\mu)\mathbf{H}_2 + \mu\mathbf{F}_{22} \end{bmatrix}\right) = 0
\end{aligned}$$

Therefore, we obtain that $Dg(\mathbf{x}^\star, \mathbf{y}^\star)^{-1}$ has an eigenvalues within $(0, 1)$ is equivalent to the event that

$$\mathbf{J}(\mu) := \begin{bmatrix} \frac{1}{\eta}(1-\mu)\mathbf{H}_1 + \mu\mathbf{F}_{11} & \sqrt{\mu}\mathbf{F}_{12} \\ \sqrt{\mu}\mathbf{F}_{21} & \frac{1}{\eta}(1-\mu)\mathbf{H}_2 + \mu\mathbf{F}_{22} \end{bmatrix}$$

is a singular matrix for certain $\mu \in (0, 1)$. Towards that end, we first observe that $\mathbf{J}(\mu)$ is a real-symmetric and continuous (with respect to μ) matrix and hence all eigenvalues of $\mathbf{J}(\mu)$ are real-valued (by the symmetry of $\mathbf{J}(\mu)$) and are continuous functions of μ ([230, Theorem 5.1]). In particular, the minimum eigenvalue $\lambda_{\min}(\mathbf{J}(\mu))$ is real-valued and continuous function of μ.

Now, we observe that

$$\begin{aligned}
\lim_{\mu \to 0^+} \mathbf{J}(\mu) &= \begin{bmatrix} \frac{1}{\eta}\mathbf{H}_1 & \\ & \frac{1}{\eta}\mathbf{H}_2 \end{bmatrix}, \\
\mathbf{J}(1) &= \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} = \begin{bmatrix} \nabla^2_{11} f(\mathbf{x}^\star, \mathbf{y}^\star) & \nabla^2_{12} f(\mathbf{x}^\star, \mathbf{y}^\star) \\ \nabla^2_{21} f(\mathbf{x}^\star, \mathbf{y}^\star) & \nabla^2_{22} f(\mathbf{x}^\star, \mathbf{y}^\star) \end{bmatrix} = \nabla^2 f(\mathbf{x}^\star, \mathbf{y}^\star)
\end{aligned}$$

First, since $\mathbf{J}(0^+)$ is positive definite (as h_1 and h_2 are strongly convex), we have

$$\lambda_{\min}(\mathbf{J}(0^+)) > 0.$$

Second, since $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a strict saddle of f, we claim that

$$\lambda_{\min}(\mathbf{J}(1)) < 0.$$

Finally, since $\lambda_{\min}(\mathbf{J}(\mu))$ is a real-valued and continuous function of μ, we get $\lambda_{\min}(\mathbf{J}(\mu)) = 0$ for some $\mu \in (0, 1)$.

This completes the argument of showing Algorithm 4 can almost surely avoid the strict saddles. Then together with the first-order convergence Theorem G.4.3, we obtain the second-order convergence of Algorithm 4. □

G.4.3 Convergence Analysis of Algorithm 5

Algorithm 12 Bregman Proximal Minimization

1: **Input:** A Bregman kernel h with $L_f\nabla^2 h(\mathbf{x}) \pm \nabla^2 f(\mathbf{x}) \succeq 0$ in the whole domain; Set $\eta \in (0, \frac{1}{L_f})$.

2: **Initialization:** $\mathbf{x}^0$

3: **Recursion:** Iteratively generate a sequence $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ via

$$\mathbf{x}^{\ell+1} = g(\mathbf{x}^\ell) := \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{1}{\eta} D_h(\mathbf{x}, \mathbf{x}^\ell) \tag{10.20}$$

G.4.3.1 First-order Convergence of Algorithm 5

Theorem G.4.5. *Under Assumptions 10.2.1– 10.2.4, Algorithm 5 with arbitrary initialization converges to a critical point of f in* (10.1)*.*

Proof. First of all, Algorithm 5 is well-defined in view of Lemma 10.2.1. Then, by Theorem 10.4.1 and the assumption that f is a KL function, it is sufficient to prove that $\{\mathbf{x}^\ell\}_{\ell\in\mathbb{N}}$ is a gradient-like descent sequence for f (see Definition 10.4.1), i.e., to show:

(C1) Sufficient decrease property: $f(\mathbf{x}^\ell) - f(\mathbf{x}^{\ell+1}) \geq \rho_1 \left\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\right\|_2^2,\ \forall\, \ell \in \mathbb{N}$ for some $\rho_1 > 0$;

(C2) Bounded gradient property: $\left\|\nabla f(\mathbf{x}^{\ell+1})\right\|_2 \leq \rho_2 \left\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\right\|_2,\ \forall\, \ell \in \mathbb{N}$ for some $\rho_2 > 0$.

Condition (C1) follows from (10.11) in Lemma 10.2.1. Condition (C2) holds because by the optimality condition

$$\nabla f(\mathbf{x}^{\ell+1}) + (\nabla h(\mathbf{x}^{\ell+1}) - \nabla h(\mathbf{x}^\ell))/\eta = \mathbf{0}, \tag{G.47}$$

we have $\|\nabla f(\mathbf{x}^{\ell+1})\|_2 = \frac{1}{\eta}\|\nabla h(\mathbf{x}^{\ell+1}) - \nabla h(\mathbf{x}^\ell)\|_2 \leq \frac{\rho_h(\mathcal{B})}{\eta}\|\mathbf{x}^{\ell+1} - \mathbf{x}^\ell\|_2$, where the inequality follows from Assumption 10.2.4, $h \in \mathcal{C}^2$, and Footnote 40. □

G.4.3.2 Second-order Convergence of Algorithm 5

Theorem G.4.6. *Under Assumptions 10.2.1– 10.2.4, Algorithm 5 with random initialization almost surely converges to a second-order stationary point of f in* (10.1)*.*

Proof. To show the second-order convergence, we define (10.20) as $\mathbf{x}^\ell = g(\mathbf{x}^{\ell-1})$ and compute the Jacobian matrix Dg. By the definition of g, we have $Dg(\mathbf{x}^\ell) = \partial\mathbf{x}^{\ell+1}/\partial\mathbf{x}^\ell$. Now we apply the implicit function theorem to (G.47) and in view of the nonsingularity of $\nabla^2 h + \eta\nabla^2 f$, we obtain that Dg is continuous and given by

$$Dg(\mathbf{x}^\ell) = \left(\nabla^2 h(\mathbf{x}^{\ell+1}) + \eta\nabla^2 f(\mathbf{x}^{\ell+1})\right)^{-1} \nabla^2 h(\mathbf{x}^\ell).$$

Noting that the above argument holds for any $\mathbf{x}^\ell \in \mathbb{R}^n$, we therefore have that $Dg(\mathbf{x})$ is continuous and given by

$$Dg(\mathbf{x}) = \left(\nabla^2 h(g(\mathbf{x})) + \eta\nabla^2 f(g(\mathbf{x}))\right)^{-1} \nabla^2 h(\mathbf{x}). \tag{G.48}$$

By Theorem 10.4.2, to show the mapping g can almost surely avoid the strict saddles, it suffices to show the following conditions:

Showing g is a $\mathcal{C}^1$ mapping. This immediately follows from the continuity of Dg in (G.48).

Showing $\det(Dg) \neq 0$ in the whole domain. Due to the positive definiteness of $\nabla^2 h$ and $\nabla^2 h \pm \eta\nabla^2 f$,

$$\det(Dg(\mathbf{x})) = \det\left(\left[\nabla^2 h(g(\mathbf{x})) + \eta\nabla^2 f(g(\mathbf{x}))\right]^{-1}\right) \det\left(\nabla^2 h(\mathbf{x})\right) > 0.$$

Showing any strict saddle of f lies in $\mathcal{A}_g$. First for any strict saddle $\mathbf{x}^\star$, we have $\mathbf{x}^{\ell+1} = \mathbf{x}^\ell = \mathbf{x}^\star$ satisfies the optimality condition (G.47), indicating $\mathbf{x}^\star$ is a fixed point, i.e., $g(\mathbf{x}^\star) = \mathbf{x}^\star$. Now plugging $g(\mathbf{x}^\star) = \mathbf{x}^\star$ to (G.48), we have

$$\begin{aligned}
Dg(\mathbf{x}^\star) =& [\nabla^2 h(\mathbf{x}^\star) + \eta\nabla^2 f(\mathbf{x}^\star)]^{-1}\nabla^2 h(\mathbf{x}^\star) \\
\sim& [\nabla^2 h(\mathbf{x}^\star) + \eta\nabla^2 f(\mathbf{x}^\star)]^{-1/2}(\nabla^2 h(\mathbf{x}^\star))[\nabla^2 h(\mathbf{x}^\star) + \eta\nabla^2 f(\mathbf{x}^\star)]^{-1/2} \\
=& \mathbf{I} - \eta[\nabla^2 h(\mathbf{x}^\star) + \eta\nabla^2 f(\mathbf{x}^\star)]^{-1/2}\nabla^2 f(\mathbf{x}^\star)[\nabla^2 h(\mathbf{x}^\star) + \eta\nabla^2 f(\mathbf{x}^\star)]^{-1/2} := \mathbf{I} - \eta\mathbf{\Phi}
\end{aligned}$$

where "$\sim$" denotes matrix-similarity. Clearly, we know $Dg(\mathbf{x}^\star)$ has an eigenvalue strictly greater than 1 since $\nabla^2 f(\mathbf{x}^\star)$ has a negative eigenvalue and is congruent to $\mathbf{\Phi}$.

Combining the above three and Theorem 10.4.2, we show that Algorithm 5 can almost surely avoid strict saddles. Finally, combining this with the first-order convergence, we obtain the second-order convergence of Algorithm 5. □

G.4.4 Convergence Analysis of Algorithm 6

Algorithm 13 Bregman Proximal Alternating Minimization

1: **Input:** A bi-Bregman kernel $h(\mathbf{x}, \mathbf{y})$ with both $L_1\nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y}) \pm \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) \succeq 0$ and $L_2\nabla^2_{\mathbf{yy}} h(\mathbf{x}, \mathbf{y}) \pm \nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \succeq 0$ in the entire domain; Set $\eta \in (0, \min(\frac{1}{L_1}, \frac{1}{L_2}))$.

2: **Initialization:** $(\mathbf{x}^0, \mathbf{y}^0)$

3: **Recursion:** Iteratively generate a sequence $\{\mathbf{x}^\ell, \mathbf{y}^\ell\}_{\ell\in\mathbb{N}}$ via

$$\begin{aligned}
\mathbf{x}^\ell =& \arg\min_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^{\ell-1}) + \frac{1}{\eta} D_h^1(\mathbf{x}, \mathbf{x}^{\ell-1}; \mathbf{y}^{\ell-1}), \\
\mathbf{y}^\ell =& \arg\min_{\mathbf{y}} f(\mathbf{x}^\ell, \mathbf{y}) + \frac{1}{\eta} D_h^2(\mathbf{y}, \mathbf{y}^{\ell-1}; \mathbf{x}^\ell)
\end{aligned} \tag{G.49}$$

G.4.4.1 First-order Convergence of Algorithm 6

Theorem G.4.7. *Under Assumptions 10.2.1– 10.2.4, Algorithm 6 with arbitrary initialization converges to a critical point of f in in* (10.12).

Proof. First of all, as a direct consequence of Lemma 10.2.2, we are guaranteed Algorithm 6 is well defined.

Proposition G.4.2. *Under Assumptions 10.2.1– 10.2.4, Algorithm 6 is well-defined.*

Now, in view of Theorem 10.4.1 and the assumption that f is KL function, it is sufficient to prove that $\{(\mathbf{x}^\ell, \mathbf{y}^\ell)\}_{\ell\in\mathbb{N}}$ is a gradient-like descent sequence for f (see Definition 10.4.1), i.e., satisfying

(C1) Sufficient decrease property;

(C2) Bounded gradient property.

Condition (C1) directly follows from Lemma 10.2.2.

To show Condition (C2), we start with the optimality condition of the first block of Algorithm 6

$$\nabla_{\mathbf{x}} h(\mathbf{x}^+, \mathbf{y}) = \nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{y}) - \eta \nabla_{\mathbf{x}} f(\mathbf{x}^+, \mathbf{y}), \tag{G.50}$$

and get that

$$\begin{aligned}\|\nabla_{\mathbf{x}} f(\mathbf{x}^+, \mathbf{y})\|_2 &= \frac{1}{\eta}\|\nabla_{\mathbf{x}} h(\mathbf{x}^+, \mathbf{y}^\star) - \nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{y})\|_2 \\ &\leq \frac{\rho_h(\mathcal{B})}{\eta}\|\mathbf{x}^+ - \mathbf{x}\|_2\end{aligned}$$

where the second line follows from Assumption 10.2.4, $h \in \mathcal{C}^2$, and Footnote 40.

Now using the same argument on $f \in \mathcal{C}^2$, we get f has a locally $\rho_f(\mathcal{B})$-Lipschitz gradient on the set $\mathcal{B}$, and therefore

$$\begin{aligned}\|\nabla_{\mathbf{x}} f(\mathbf{x}^+, \mathbf{y}^+)\|_2 &\leq \|\nabla_{\mathbf{x}} f(\mathbf{x}^+, \mathbf{y}^+) - \nabla_{\mathbf{x}} f(\mathbf{x}^+, \mathbf{y})\|_2 + \|\nabla_{\mathbf{x}} f(\mathbf{x}^+, \mathbf{y})\|_2 \\ &\leq \rho_f(\mathcal{B})\|\mathbf{y}^+ - \mathbf{y}\|_2 + \frac{\rho_h(\mathcal{B})}{\eta}\|\mathbf{x}^+ - \mathbf{x}\|_2 \\ &\leq \left(\rho_f(\mathcal{B}) + \frac{\rho_h(\mathcal{B})}{\eta}\right)\|(\mathbf{x}^+, \mathbf{y}^+) - (\mathbf{x}, \mathbf{y})\|_2\end{aligned}$$

Using a similar analysis to the optimality condition for the second-block of Algorithm 6

$$\nabla_{\mathbf{y}} h(\mathbf{x}^+, \mathbf{y}^+) = \nabla_{\mathbf{y}} h(\mathbf{x}^+, \mathbf{y}) - \eta \nabla_{\mathbf{y}} f(\mathbf{x}^+, \mathbf{y}^+),$$

we can get

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}^+, \mathbf{y}^+)\|_2 \leq \frac{\rho_h(\mathcal{B})}{\eta} \|\mathbf{y}^+ - \mathbf{y}\|_2$$

Therefore, combining all, we get

$$\begin{aligned}\|\nabla f(\mathbf{x}^+, \mathbf{y}^+)\|_2 \leq & \|\nabla_{\mathbf{x}} f(\mathbf{x}^+, \mathbf{y}^+)\|_2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}^+, \mathbf{y}^+)\|_2 \\ \leq & \left(\frac{2\rho_h(\mathcal{B})}{\eta} + \rho_f(\mathcal{B})\right) \|(\mathbf{x}^+, \mathbf{y}^+) - (\mathbf{x}, \mathbf{y})\|_2\end{aligned}$$

□

G.4.4.2 Second-order Convergence of Algorithm 6

Theorem G.4.8. *Under Assumptions 10.2.1– 10.2.4, Algorithm 6 with random initialization almost surely converges to a second-order stationary point of f in* (10.12).

Proof. Following from (G.49), we denote

$$\begin{aligned}(\mathbf{x}^+, \mathbf{y}) &= g_1(\mathbf{x}, \mathbf{y}) \\ (\mathbf{x}, \mathbf{y}^+) &= g_2(\mathbf{x}, \mathbf{y})\end{aligned} \tag{G.51}$$

The mapping g_1, g_2 are well-defined in the whole domain $\mathbb{R}^n \times \mathbb{R}^m$, in view of strong convexity and coercivity of the objective function in (G.49). Then Algorithm 6 can be viewed as iteratively performing the following composite mapping for $\ell = 1, 2, \ldots$

$$(\mathbf{x}^\ell, \mathbf{y}^\ell) = g(\mathbf{x}^{\ell-1}, \mathbf{y}^{\ell-1}) \tag{G.52}$$

with the mapping g defined as the composite mapping of g_1, g_2:

$$g := g_2 \circ g_1. \tag{G.53}$$

To compute the the Jacobian matrix Dg, We first compute Dg_1 and Dg_2. Then we will use the chain rule to get Dg.

Now we compute $Dg_1(\mathbf{x}, \mathbf{y})$ from $(\mathbf{x}^+, \mathbf{y}) = g_1(\mathbf{x}, \mathbf{y})$. First of all,

$$Dg_1(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \frac{\partial \mathbf{x}^+}{\partial \mathbf{x}^\top} & \frac{\partial \mathbf{x}^+}{\partial \mathbf{y}^\top} \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix}$$

Then recall that the first-order optimality condition (G.50) for the first block of (6) is given by

$$\nabla_{\mathbf{x}} h(\mathbf{x}^+, \mathbf{y}) = \nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{y}) - \eta \nabla_{\mathbf{x}} f(\mathbf{x}^+, \mathbf{y})$$

Now apply the Implicit function theorem to (G.50)

$$\left(\nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y}) + \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x}^+, \mathbf{y})\right) \frac{\partial \mathbf{x}^+}{\partial \mathbf{x}^\top} = \nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y})$$
$$\left(\nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y}) + \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x}^+, \mathbf{y})\right) \frac{\partial \mathbf{x}^+}{\partial \mathbf{y}^\top} = \nabla^2_{\mathbf{xy}} h(\mathbf{x}, \mathbf{y}) - \nabla^2_{\mathbf{xy}} h(\mathbf{x}^+, \mathbf{y}) - \eta \nabla^2_{\mathbf{xy}} f(\mathbf{x}^+, \mathbf{y})$$

implying (since $\nabla^2 h_1$ is positive definite in the whole domain) that the following Jacobians are continuous:

$$\frac{\partial \mathbf{x}^+}{\partial \mathbf{x}^\top} = \left(\nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y}) + \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x}^+, \mathbf{y})\right)^{-1} \nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y})$$
$$\frac{\partial \mathbf{x}^+}{\partial \mathbf{y}^\top} = \left(\nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y}) + \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x}^+, \mathbf{y})\right)^{-1} \left(\nabla^2_{\mathbf{xy}} h(\mathbf{x}, \mathbf{y}) - \nabla^2_{\mathbf{xy}} h(\mathbf{x}^+, \mathbf{y}) - \eta \nabla^2_{\mathbf{xy}} f(\mathbf{x}^+, \mathbf{y})\right)$$

Therefore, Dg_1 is continuous and is given by

$$\begin{aligned} Dg_1(\mathbf{x}, \mathbf{y}) &= \begin{bmatrix} \frac{\partial \mathbf{x}^+}{\partial \mathbf{x}^\top} & \frac{\partial \mathbf{x}^+}{\partial \mathbf{y}^\top} \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix} \\ &= \begin{bmatrix} \left(\nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y}) + \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x}^+, \mathbf{y})\right)^{-1} & \\ & \mathbf{I}_m \end{bmatrix} \\ &\quad \begin{bmatrix} \nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y}) & \nabla^2_{\mathbf{xy}} h(\mathbf{x}, \mathbf{y}) - \nabla^2_{\mathbf{xy}} h(\mathbf{x}^+, \mathbf{y}) - \eta \nabla^2_{\mathbf{xy}} f(\mathbf{x}^+, \mathbf{y}) \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix} \end{aligned} \tag{G.54}$$

Similarly, we have Dg_2 is continuous and given by

$$\begin{aligned} Dg_2(\mathbf{x}, \mathbf{y}) &= \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \frac{\partial \mathbf{y}^+}{\partial \mathbf{x}^\top} & \frac{\partial \mathbf{y}^+}{\partial \mathbf{y}^\top} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I}_n & \\ & \left(\nabla^2 h_2(\mathbf{y}^+) + \eta \nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}^+)\right)^{-1} \end{bmatrix} \\ &\quad \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \nabla^2_{\mathbf{yx}} h(\mathbf{x}, \mathbf{y}) - \nabla^2_{\mathbf{yx}} h(\mathbf{x}, \mathbf{y}^+) - \eta \nabla^2_{\mathbf{yx}} f(\mathbf{x}, \mathbf{y}^+) & \nabla^2_{\mathbf{yy}} h(\mathbf{x}, \mathbf{y}) \end{bmatrix} \end{aligned} \tag{G.55}$$

Finally, combining (G.54) and (G.55), we get the expression of Dg:

$$Dg(\mathbf{x}, \mathbf{y}) = Dg_2(g_1(\mathbf{x}, \mathbf{y})) Dg_1(\mathbf{x}, \mathbf{y}). \tag{G.56}$$

Further, since continuity is preserved by product and composite operation, we get Dg is continuous and hence $g \in \mathcal{C}^1$.

By Theorem 10.4.2, to show that the mapping g can almost surely avoid the strict saddles, it suffices to show the following conditions:

1. g is $\mathcal{C}^1$ mapping;
2. $\det(Dg) \neq 0$ in the whole domain;
3. Any strict saddle of f is an unstable fixed point of g.

Showing g is $\mathcal{C}^1$ mapping. This is because Dg in (G.56) is continuous by the implicit function theorem.

Showing $\det(Dg) \neq 0$ in the whole domain. Because

$$Dg(\mathbf{x}, \mathbf{u}) = Dg_2(g_1(\mathbf{x}, \mathbf{y}))Dg_1(\mathbf{x}, \mathbf{y})$$

with Dg_1 and Dg_2 being square matrices, it suffices to show the global non-singularity of both Dg_1 and Dg_2. Since Dg_1 is a block upper-triangular matrix in view of (G.54), it suffices to show both of its diagonal block matrices are nonsingular. The first diagonal block is $\left(\nabla^2_{\mathbf{xx}} h(\mathbf{x}^+, \mathbf{y}) + \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x}^+, \mathbf{y})\right)^{-1} \nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y})$ in (G.54), which is nonsingular in the whole domain because of h is strongly bi-convex (implying $\nabla^2_{\mathbf{xx}} h$ is positive definite in the whole domain and hence nonsingular in the whole domain) and (f, h) satisfies (L_1, L_2)-bi-adaptive Lipschitz gradient condition (implying $\nabla^2_{\mathbf{xx}} h(\mathbf{x}, \mathbf{y}) + \eta \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y})$ is positive definite in the whole domain for any $\eta < \frac{1}{L_1}$ and hence nonsingular in the whole domain). Therefore, we obtain that Dg_1 is nonsingular in the whole domain. Use a similar analysis and in view of (G.55), and we can show that $Dg_2(\mathbf{x}, \mathbf{y})$ is nonsingular in the whole domain. This shows that $\det(Dg) \neq 0$ in the whole domain.

Showing any strict saddle of f lies in $\mathcal{A}_g$. First of all, we show that for any strict saddle $(\mathbf{x}^\star, \mathbf{y}^\star)$ of f, we have

$$\begin{aligned} g_1(\mathbf{x}^\star, \mathbf{y}^\star) &= (\mathbf{x}^\star, \mathbf{y}^\star), \\ g_2(\mathbf{x}^\star, \mathbf{y}^\star) &= (\mathbf{x}^\star, \mathbf{y}^\star). \end{aligned}$$

Then this implies $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a fixed point the mapping $g = g_2 \circ g_1$, i.e., $g(\mathbf{x}^\star, \mathbf{y}^\star) = (\mathbf{x}^\star, \mathbf{y}^\star)$. Clearly $\mathbf{x} = \mathbf{x}^\star, \mathbf{x}^+ = \mathbf{x}^\star, \mathbf{y} = \mathbf{y}^\star$ satisfies the first-order optimality condition (G.50). Combining the well-definedness of g_1 by Proposition G.4.2, this implies that

$$g_1(\mathbf{x}^\star, \mathbf{y}^\star) = (\mathbf{x}^\star, \mathbf{y}^\star).$$

The same analysis can be used to show that

$$g_2(\mathbf{x}^\star, \mathbf{y}^\star) = (\mathbf{x}^\star, \mathbf{y}^\star).$$

This completes the proof of showing $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a fixed point of g.

It remains to show that the Jacobian matrix $Dg(\mathbf{x}^\star, \mathbf{y}^\star)$ has an eigenvalues with magnitude greater than 1. To simplify notations, we make the following notations;

$$\begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} := \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\mathbf{x}^\star, \mathbf{y}^\star) & \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star, \mathbf{y}^\star) \\ \nabla^2_{\mathbf{yx}} f(\mathbf{x}^\star, \mathbf{y}^\star) & \nabla^2_{\mathbf{yy}} f(\mathbf{x}^\star, \mathbf{y}^\star) \end{bmatrix}$$

and

$$\mathbf{H}_1 := \nabla^2_{\mathbf{xx}} h(\mathbf{x}^\star, \mathbf{y}^\star)$$
$$\mathbf{H}_2 := \nabla^2_{\mathbf{yy}} h(\mathbf{x}^\star, \mathbf{y}^\star).$$

Now we are ready to compute $Dg(\mathbf{x}^\star, \mathbf{y}^\star)$ by plugging

$$(\mathbf{x}^+, \mathbf{y}^+) = (\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\star, \mathbf{y}^\star)$$

to (G.56) and using the above notations:

$$\begin{aligned}
Dg(\mathbf{x}^\star, \mathbf{y}^\star) =& Dg_2(\mathbf{x}^\star, \mathbf{y}^\star) Dg_1(\mathbf{x}^\star, \mathbf{y}^\star) \\
=& \left(\begin{bmatrix} \mathbf{I}_n & \\ & (\mathbf{H}_2 + \eta \mathbf{F}_{22})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ -\eta \mathbf{F}_{21} & \mathbf{H}_2 \end{bmatrix} \right) \left(\begin{bmatrix} (\mathbf{H}_1 + \eta \mathbf{F}_{11})^{-1} & \\ & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{H}_1 & -\eta \mathbf{F}_{12} \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix} \right) \\
=& \begin{bmatrix} (\mathbf{H}_1 + \eta \mathbf{F}_{11})^{-1} & \mathbf{0} \\ (\mathbf{H}_1 + \eta \mathbf{F}_{11})^{-1}(-\eta \mathbf{F}_{21})(\mathbf{H}_2 + \eta \mathbf{F}_{22})^{-1} & (\mathbf{H}_2 + \eta \mathbf{F}_{22})^{-1} \mathbf{H}_2 \end{bmatrix} \begin{bmatrix} \mathbf{H}_1 & -\eta \mathbf{F}_{12} \\ & \mathbf{I}_m \end{bmatrix} \\
=& \begin{bmatrix} (\mathbf{H}_1 + \eta \mathbf{F}_{11})^{-1} & \mathbf{0} \\ (\mathbf{H}_1 + \eta \mathbf{F}_{11})^{-1}(-\eta \mathbf{F}_{21})(\mathbf{H}_2 + \eta \mathbf{F}_{22})^{-1} & (\mathbf{H}_2 + \eta \mathbf{F}_{22})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \\ & \mathbf{H}_2 \end{bmatrix} \begin{bmatrix} \mathbf{H}_1 & -\eta \mathbf{F}_{12} \\ & \mathbf{I}_m \end{bmatrix} \\
=& \begin{bmatrix} \mathbf{H}_1 + \eta \mathbf{F}_{11} & \mathbf{0} \\ \eta \mathbf{F}_{21} & \mathbf{H}_2 + \eta \mathbf{F}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{H}_1 & -\eta \mathbf{F}_{12} \\ & \mathbf{H}_2 \end{bmatrix}
\end{aligned}$$

Second, we transform the problem of showing that $Dg(\mathbf{x}^\star, \mathbf{y}^\star)$ has an eigenvalue of magnitude greater than 1 as the problem of showing that

$$\det(Dg(\mathbf{x}^\star, \mathbf{y}^\star) - \mu \mathbf{I}) = 0$$

for some μ of magnitude greater than 1. Using the properties of $\det(\cdot)$, we further have

$$\begin{aligned}
& \det(Dg(\mathbf{x}^\star, \mathbf{y}^\star) - \mu \mathbf{I}) = 0 \\
\iff & \det\left(\begin{bmatrix} \mathbf{H}_1 + \eta \mathbf{F}_{11} & \mathbf{0} \\ \eta \mathbf{F}_{21} & \mathbf{H}_2 + \eta \mathbf{F}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{H}_1 & -\eta \mathbf{F}_{12} \\ & \mathbf{H}_2 \end{bmatrix} - \mu \mathbf{I} \right) = 0 \\
\iff & \det\left(\begin{bmatrix} \mathbf{H}_1 & -\eta \mathbf{F}_{12} \\ & \mathbf{H}_2 \end{bmatrix} - \mu \begin{bmatrix} \mathbf{H}_1 + \eta \mathbf{F}_{11} & \mathbf{0} \\ \eta \mathbf{F}_{21} & \mathbf{H}_2 + \eta \mathbf{F}_{22} \end{bmatrix} \right) = 0 \\
\iff & \det\left(\begin{bmatrix} (1-\mu)\mathbf{H}_1 - \mu\eta \mathbf{F}_{11} & -\eta \mathbf{F}_{12} \\ -\mu \eta \mathbf{F}_{21} & (1-\mu)\mathbf{H}_2 - \mu \eta \mathbf{F}_{22} \end{bmatrix} \right) = 0 \\
\iff & \det\left(\begin{bmatrix} (\mu-1)\mathbf{H}_1 + \mu\eta \mathbf{F}_{11} & \eta \mathbf{F}_{12} \\ \mu \eta \mathbf{F}_{21} & (\mu-1)\mathbf{H}_2 + \mu \eta \mathbf{F}_{22} \end{bmatrix} \right) = 0 \\
\iff & \det\left(\begin{bmatrix} \mathbf{I}_n & \\ & \sqrt{\mu}\mathbf{I}_m \end{bmatrix} \begin{bmatrix} (\mu-1)\mathbf{H}_1 + \mu\eta \mathbf{F}_{11} & \sqrt{\mu}\eta \mathbf{F}_{12} \\ \sqrt{\mu}\eta \mathbf{F}_{21} & (\mu-1)\mathbf{H}_2 + \mu \eta \mathbf{F}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & \\ & \sqrt{\mu}\mathbf{I}_m \end{bmatrix}^{-1} \right) = 0 \\
\iff & \det\left(\begin{bmatrix} \mathbf{I}_n & \\ & \sqrt{\mu}\mathbf{I}_m \end{bmatrix} \right) \det\left(\begin{bmatrix} (\mu-1)\mathbf{H}_1 + \mu\eta \mathbf{F}_{11} & \sqrt{\mu}\eta \mathbf{F}_{12} \\ \sqrt{\mu}\eta \mathbf{F}_{21} & (\mu-1)\mathbf{H}_2 + \mu \eta \mathbf{F}_{22} \end{bmatrix} \right) \det\left(\begin{bmatrix} \mathbf{I}_n & \\ & \sqrt{\mu}\mathbf{I}_m \end{bmatrix}^{-1} \right) = 0 \\
\iff & \det\left(\begin{bmatrix} (\mu-1)\mathbf{H}_1 + \mu\eta \mathbf{F}_{11} & \sqrt{\mu}\eta \mathbf{F}_{12} \\ \sqrt{\mu}\eta \mathbf{F}_{21} & (\mu-1)\mathbf{H}_2 + \mu \eta \mathbf{F}_{22} \end{bmatrix} \right) = 0
\end{aligned}$$

Therefore, the problem reduces to showing that

$$\mathbf{J}(\mu) := \begin{bmatrix} (\mu-1)\mathbf{H}_1 + \mu\eta\mathbf{F}_{11} & \sqrt{\mu}\eta\mathbf{F}_{12} \\ \sqrt{\mu}\eta\mathbf{F}_{21} & (\mu-1)\mathbf{H}_2 + \mu\eta\mathbf{F}_{22} \end{bmatrix}$$

is a singular matrix for some $\mu > 1$. Note that $\mathbf{J}(\mu)$ is a symmetric and continuous (with respect to μ) matrix and hence all the eigenvalues are real-valued (by symmetric structure of $\mathbf{J}(\mu)$) and continuous functions of μ (by [230, Theorem 5.1]). In particular, the minimum eigenvalue $\lambda_{\min}(\mathbf{J}(\mu)))$ is also a real-valued and continuous function of μ.

Now we observe $\mathbf{J}(\mu)$ in two special cases:

$$\begin{aligned} \mathbf{J}(1) &= \eta \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} = \eta \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\mathbf{x}^\star, \mathbf{y}^\star) & \nabla^2_{\mathbf{xy}} f(\mathbf{x}^\star, \mathbf{y}^\star) \\ \nabla^2_{\mathbf{yx}} f(\mathbf{x}^\star, \mathbf{y}^\star) & \nabla^2_{\mathbf{yy}} f(\mathbf{x}^\star, \mathbf{y}^\star) \end{bmatrix} = \eta \nabla^2 f(\mathbf{x}^\star, \mathbf{y}^\star), \\ \lim_{\mu\to\infty} \frac{\mathbf{J}(\mu)}{\mu} &= \begin{bmatrix} \mathbf{H}_1 + \eta\mathbf{F}_{11} & \\ & \mathbf{H}_2 + \eta\mathbf{F}_{22} \end{bmatrix} \end{aligned}$$

First, since $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a strict saddle of f, we have

$$\lambda_{\min}(\mathbf{J}(1)) < 0.$$

Second, by the assumption that both $\mathbf{H}_1 \pm \eta\mathbf{F}_{11}$ and $\mathbf{H}_2 \pm \eta\mathbf{F}_{22}$ are positive definite matrices, we have

$$\lambda_{\min}(\mathbf{J}(N)) > 0$$

for some sufficiently large number N. Finally, since $\lambda_{\min}(\mathbf{J}(\mu)))$ is a real-valued and continuous function of μ, we conclude that there must be $\mu \in (1, N)$ such that $\lambda_{\min}(\mathbf{J}(\mu)) = 0$ for some sufficiently large number $N > 1$.

Therefore, we have shown that Algorithm 6 can almost surely avoid the strict saddles. Combining this with the first-order convergence Theorem G.4.7, we obtain the second-order convergence of Algorithm 6. □

APPENDIX H

APPENDICES FOR CHAPTER 11

H.1 Proof of Theorem 11.5.2

Definition H.1.1 (KL property)**.** *[212, 241] We say a proper semi-continuous function $h(\mathbf{u})$ satisfies Kurdyka-Łojasiewicz property, if for any limiting critical point $\overline{\mathbf{u}}$ of $h(\mathbf{u})$, there exist $\delta > 0,\ \theta \in [0,1),\ C > 0$, s.t.*

$$|h(\mathbf{u}) - h(\overline{\mathbf{u}})|^{\theta} \leq C\,\mathrm{dist}(0, \partial h(\mathbf{u})),\ \ \forall\, \mathbf{u} \in \mathcal{B}(\overline{\mathbf{u}}, \delta)$$

where $\partial h(\mathbf{u}))$ denotes the subdifferential of h at μ. In particular, when $h(\cdot)$ is differentiable, we further have $\partial h(\mathbf{u})) = \{\nabla h(\mathbf{u})\}$ and hence the KL property becomes

$$|h(\mathbf{u}) - h(\overline{\mathbf{u}})|^{\theta} \leq C\|\nabla h(\mathbf{u})\|_2,\ \ \forall\, \mathbf{u} \in \mathcal{B}(\overline{\mathbf{u}}, \delta)$$

We mention that the above KL property (also known as KL inequality) states the regularity of $h(\mathbf{u})$ around its critical point $\mathbf{u}$ and the KL inequality trivially holds at non-critical point. There are a very large set of functions satisfying the KL inequality. In particular [213, Theorem 5.1], any proper lower semi-continuous function satisfies the KL property once its function graph is a semi-algebraic set, i.e., is a subset of $\mathbb{R}^n$ defined by a finite sequence of polynomial equations. Therefore, a very large set of functions should satisfy the KL inequality, since the semi-algebraic property functions are sufficiently general, including but never limited to any polynomials, any norm, quasi norm, ℓ_0 norm, smooth manifold, etc. For more discussions and examples, see [213, 239]. Clearly, the objective function $f(\mathbf{U}, \mathbf{V}, \mathbf{W})$ is semi-algebraic as it is a polynomial function.

Lemma H.1.1 (Uniform KL property)**.** *The objective function in eq.* (11.7) *satisfies the KL property. Further there exist $\delta_0 > 0,\ \theta_{KL} \in [0,1),\ C_{KL} > 0$ such that as long as $\mathrm{dist}((\mathbf{U}, \mathbf{V}, \mathbf{W}), \mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)) \leq \delta_0$, we have*

$$\left|f(\mathbf{U}, \mathbf{V}, \mathbf{W}) - \overline{f}\right|^{\theta_{KL}} \leq C_{KL}\|\nabla f(\mathbf{U}, \mathbf{V}, \mathbf{W})\|_F \tag{H.1}$$

with $\overline{f}$ being limiting function value defined in Part (i) of Theorem 11.5.1.

Proof. First recognize the union $\bigcup_i \mathcal{B}((\overline{\mathbf{U}}_i, \overline{\mathbf{V}}_i, \overline{\mathbf{W}}_i), \delta_i)$ forms an open cover of $\mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$ with $(\overline{\mathbf{U}}_i, \overline{\mathbf{V}}_i, \overline{\mathbf{W}}_i)$ representing all points in $\mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$ and δ_i to be chosen so that the the following KL property of f at $(\overline{\mathbf{U}}_i, \overline{\mathbf{V}}_i, \overline{\mathbf{W}}_i) \in \mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$ holds:

$$\left|f(\mathbf{U}, \mathbf{V}, \mathbf{W}) - \overline{f}\right|^{\theta_i} \leq C_i\|\nabla f(\mathbf{U}, \mathbf{V}, \mathbf{W})\|_F,\ \forall (\mathbf{U}, \mathbf{V}, \mathbf{W}) \in \mathcal{B}((\overline{\mathbf{U}}_i, \overline{\mathbf{V}}_i, \overline{\mathbf{W}}_i), \delta_i)$$

where we have used all $f(\overline{\mathbf{U}}_i, \overline{\mathbf{V}}_i, \overline{\mathbf{W}}_i) = \overline{f}$ by Part (iii) of Theorem 11.5.1. Then due to the compactness of the set $\mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$ (from Part (iv) of Theorem 11.5.1), it has a finite subcover, that is, $\bigcup_{i=1}^{p} \mathcal{B}((\overline{\mathbf{U}}_{k_i}, \overline{\mathbf{V}}_{k_i}, \overline{\mathbf{W}}_{k_i}), \delta_{k_i})$

for some positive integer p. Now combining all, we have for all $(\mathbf{U},\mathbf{V},\mathbf{W}) \in \bigcup_{i=1}^{p} \mathcal{B}((\overline{\mathbf{U}}_{k_i},\overline{\mathbf{V}}_{k_i},\overline{\mathbf{W}}_{k_i}),\delta_{k_i})$,

$$\left|f(\mathbf{U},\mathbf{V},\mathbf{W})-\overline{f}\right|^{\theta_{KL}} \leq C_{KL}\|\nabla f(\mathbf{U},\mathbf{V},\mathbf{W})\|_F \tag{H.2}$$

with $\theta_{KL}=\max_{i=1}^{p}\{\theta_{k_i}\}$ and $C_{KL}=\max_{i=1}^{p}\{C_{k_i}\}$. Finally, since $\bigcup_{i=1}^{p} \mathcal{B}((\overline{\mathbf{U}}_{k_i},\overline{\mathbf{V}}_{k_i},\overline{\mathbf{W}}_{k_i}),\delta_{k_i})$ is an open cover of $\mathcal{L}(\mathbf{U}_0,\mathbf{V}_0,\mathbf{W}_0)$, there exists a sufficiently small number δ_0 so that

$$\{(\mathbf{U},\mathbf{V},\mathbf{W}) : \operatorname{dist}((\mathbf{U},\mathbf{V},\mathbf{W}),\mathcal{L}(\mathbf{U}_0,\mathbf{V}_0,\mathbf{W}_0)) \leq \delta_0\} \subset \bigcup_{i=1}^{p} \mathcal{B}((\overline{\mathbf{U}}_{k_i},\overline{\mathbf{V}}_{k_i},\overline{\mathbf{W}}_{k_i}),\delta_{k_i}).$$

Therefore, eq. (H.2) holds for any $(\mathbf{U},\mathbf{V},\mathbf{W})$ in the δ_0-neighborhood of $\mathcal{L}(\mathbf{U}_0,\mathbf{V}_0,\mathbf{W}_0)$. □

Proof of Theorem 11.5.2. First of all, in view of that

$$\lim_{k\to\infty} \operatorname{dist}((\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k),\mathcal{L}(\mathbf{U}_0,\mathbf{V}_0,\mathbf{W}_0)) = 0$$

and the definition of the convergence, there exits a positive integer k_0 so that $\operatorname{dist}((\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k),\mathcal{L}(\mathbf{U}_0,\mathbf{V}_0,\mathbf{W}_0)) \leq \delta_0$ for all $k \geq k_0$. Now using Lemma H.1.1, we have that

$$\left|f(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)-\overline{f}\right|^{\theta_{KL}} \leq C_{KL}\|\nabla f(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)\|_F, \quad \forall k \geq k_0. \tag{H.3}$$

In the following we will restrict our iterates $\{(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)\}_{k\in\mathbb{N}}$ to $k \geq k_0$. The remaining analysis is discussed case by case.

Case I. $f(\mathbf{U}_N,\mathbf{V}_N,\mathbf{W}_N) = \overline{f}$ for some finite $N > 0$. Then by Part (i) of Theorem 11.5.1, we immediately have $f(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k) = f(\mathbf{U}_{k+1},\mathbf{V}_{k+1},\mathbf{W}_{k+1}) = \overline{f}$ for all $k \geq N$. Therefore, by Lemma 11.4.7, we get that $(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k) = (\mathbf{U}_N,\mathbf{V}_N,\mathbf{W}_N), \forall k \geq N$, hence the sequence $\{(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)\}_{k\in\mathbb{N}}$ converges to the point $(\mathbf{U}_N,\mathbf{V}_N,\mathbf{W}_N)$ in N steps. By knowing that any limit point of $\{(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)\}_{k\in\mathbb{N}}$ is a critical point of f by Part (iv) of Theorem 11.5.1, we therefore have that in this case Algorithm 7 converges to a critical point of f in a finite number of steps.

Case II. $(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k) > \overline{f}$ for any finite k. The key is to rely on the following inequality (which can be easily obtained by using Jensen's inequality to the concave function $h(x) = x^{1-\theta}$ for $\theta \in [0,1)$ with domain $x > 0$):

$$x_2^{1-\theta} - x_1^{1-\theta} \geq (1-\theta)\frac{x_2 - x_1}{x_2^{\theta}} \;\; \forall x_1 > 0, x_2 > 0 \tag{H.4}$$

Setting $x_2 = f(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k) - \overline{f} > 0$ and $x_1 = f(\mathbf{U}_{k+1},\mathbf{V}_{k+1},\mathbf{W}_{k+1}) - \overline{f} > 0$ in eq. (H.4):

$$(f(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)-\overline{f})^{1-\theta} - (f(\mathbf{U}_{k+1},\mathbf{V}_{k+1},\mathbf{W}_{k+1})-\overline{f})^{1-\theta} \geq (1-\theta)\frac{f(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k) - f(\mathbf{U}_{k+1},\mathbf{V}_{k+1},\mathbf{W}_{k+1})}{(f(\mathbf{U}_k,\mathbf{V}_k,\mathbf{W}_k)-\overline{f})^{\theta}} \tag{H.5}$$

In the subsequent argument, we choose $\theta = \theta_{KL}$ in (H.5). From eq. (H.1) and Lemma 11.4.7, we can further lower bound $f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - f(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}, \mathbf{W}_{k+1})$ and upper bound $(f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - \overline{f})^{\theta}_{KL}$ to obtain from eq. (H.5) that (denote $\Gamma_1 := 1 - \theta_{KL}, \Gamma_2 := \frac{\lambda(1-\theta_{KL})}{C_{KL}}$):

$$
\begin{aligned}
&(f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - \overline{f})^{\Gamma_1} - (f(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}, \mathbf{W}_{k+1}) - \overline{f})^{\Gamma_1} \\
&\geq \Gamma_2 \frac{\|(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}, \mathbf{W}_{k+1}) - (\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F^2}{\|\nabla f(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F} \\
&\geq \frac{\Gamma_2}{\sqrt{2}\mathcal{L}_g} \frac{\|(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}, \mathbf{W}_{k+1}) - (\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F^2}{\|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})\|_F} \\
&= \frac{\Gamma_2}{\sqrt{2}\mathcal{L}_g} \Bigg(\frac{\|(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}, \mathbf{W}_{k+1}) - (\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F^2}{\|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})\|_F} \\
&\qquad + \|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})\|_F - \|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})\|_F \Bigg) \\
&= \frac{\Gamma_2}{\sqrt{2}\mathcal{L}_g} \big(2\|(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}, \mathbf{W}_{k+1}) - (\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F - \|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\mathbf{U}_{k-1}, \mathbf{V}_{k-1}, \mathbf{W}_{k-1})\|_F \big)
\end{aligned}
$$

Repeating the above inequality and summing up them from k_0 to N, we get

$$
\begin{aligned}
&(f(\mathbf{U}_{k_0}, \mathbf{V}_{k_0}, \mathbf{W}_{k_0}) - \overline{f})^{\Gamma_1} - (f(\mathbf{U}_{N+1}, \mathbf{V}_{N+1}, \mathbf{W}_{N+1}) - \overline{f})^{\Gamma_1} \\
&\geq \frac{\Gamma_2}{\sqrt{2}\mathcal{L}_g} \Bigg(\|(\mathbf{U}_{N+1}, \mathbf{V}_{N+1}, \mathbf{W}_{N+1}) - (\mathbf{U}_N, \mathbf{V}_N, \mathbf{W}_N)\|_F \\
&\qquad - \|(\mathbf{U}_{k_0}, \mathbf{V}_{k_0}, \mathbf{W}_{k_0}) - (\mathbf{U}_{k_0-1}, \mathbf{V}_{k_0-1}, \mathbf{W}_{k_0-1})\|_F + \sum_{k=k_0}^{N} (\|(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}, \mathbf{W}_{k+1}) - (\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F) \Bigg)
\end{aligned}
$$

Letting N go to infinity and since $\lim_{N\to\infty} f(\mathbf{U}_{N+1}, \mathbf{V}_{N+1}, \mathbf{W}_{N+1}) = \overline{f}$ by Part (i) of Theorem 11.5.1, we have

$$
\begin{aligned}
&\lim_{N\to\infty} \sum_{k=k_0}^{N} \|(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}, \mathbf{W}_{k+1}) - (\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F \\
&\leq \|(\mathbf{U}_{k_0}, \mathbf{V}_{k_0}, \mathbf{W}_{k_0}) - (\mathbf{U}_{k_0-1}, \mathbf{V}_{k_0-1}, \mathbf{W}_{k_0-1})\|_F + \frac{\Gamma_2}{\sqrt{2}\mathcal{L}_g} (f(\mathbf{U}_{k_0}, \mathbf{V}_{k_0}, \mathbf{W}_{k_0}) - \overline{f})^{1-\theta_{KL}} \\
&\leq \|(\mathbf{U}_{k_0}, \mathbf{V}_{k_0}, \mathbf{W}_{k_0}) - (\mathbf{U}_{k_0-1}, \mathbf{V}_{k_0-1}, \mathbf{W}_{k_0-1})\|_F \\
&+ \frac{\Gamma_2}{\sqrt{2}\mathcal{L}_g} \|(\mathbf{U}_{k_0}, \mathbf{V}_{k_0}, \mathbf{W}_{k_0}) - (\mathbf{U}_{k_0-1}, \mathbf{V}_{k_0-1}, \mathbf{W}_{k_0-1})\|_F^{\frac{1-\theta_{KL}}{\theta_{KL}}}
\end{aligned} \tag{H.6}
$$

where in the last line we have used eq. (11.10) and eq. (H.1). Now we observe that the last line of eq. (H.6) is finite, which shows that the sequences $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}_{k\in\mathbb{N}}$ is Cauchy and hence is a convergent sequence. Then using the same arguments as Case I, we know that the unique limit point of this Cauchy sequence is also a critical point of f.

Convergence rate. We have showed that $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}$ is convergent to a unique critical point $(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}})$. In other words, the limit point set $\mathcal{L}(\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0)$ is a singleton containing this unique critical point $(\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}})$. Now, we are ready to further bound the convergence rate of the process $(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) \to (\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}})$. The key is to

utilizing the finite length inequality eq. (H.6):

$$\sum_{k=k_0}^{\infty} \|(\mathbf{U}_{k+1}, \mathbf{V}_{k+1}, \mathbf{W}_{k+1}) - (\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F \leq \|(\mathbf{U}_{k_0}, \mathbf{V}_{k_0}, \mathbf{W}_{k_0}) - (\mathbf{U}_{k_0-1}, \mathbf{V}_{k_0-1}, \mathbf{W}_{k_0-1})\|_F + \alpha \|(\mathbf{U}_{k_0}, \mathbf{V}_{k_0}, \mathbf{W}_{k_0}) - (\mathbf{U}_{k_0-1}, \mathbf{V}_{k_0-1}, \mathbf{W}_{k_0-1})\|_F^{\frac{1-\theta_{KL}}{\theta_{KL}}}$$

with $\alpha := \frac{(C_{KL}\sqrt{2}\mathcal{L}_g)^{\frac{1}{\theta_{KL}}}}{\lambda(1-\theta_{KL})}$.

We divide the following discussion into two cases based on the value of the KL exponent θ_{KL}.

Case I: $\theta_{KL} \in [0, \frac{1}{2}]$**.** Since $\theta_{KL} \in [0, \frac{1}{2}]$, we have $\frac{1-\theta_{KL}}{\theta_{KL}} \geq 1$. For simplifying notations, define $P_k = \sum_{i=k}^{\infty} \|(\mathbf{U}_{i+1}, \mathbf{V}_{i+1}, \mathbf{W}_{i+1}) - (\mathbf{U}_i, \mathbf{V}_i, \mathbf{W}_i)\|_F$. From eq. (H.6), we know that

$$P_{k_0} \leq P_{k_0-1} - P_{k_0} + \alpha[P_{k_0-1} - P_{k_0}]^{\frac{1-\theta_{KL}}{\theta_{KL}}} \tag{H.7}$$

Since by Part (ii) of Theorem 11.5.1 $P_{k-1} - P_k \to 0$ as $k \to \infty$, there exists a positive integer k_1 such that $P_{k-1} - P_k < 1$, for all $k \geq k_1$. Then combining (H.7) and the fact $\frac{1-\theta_{KL}}{\theta_{KL}} \geq 1$, we have

$$P_k \leq (1+\alpha)(P_{k-1} - P_k), \ \forall k \geq \overline{k}$$

with $\overline{k} := \max\{k_0, k_1\}$, which further gives that

$$P_k \leq \frac{1+\alpha}{2+\alpha} P_{k-1}, \ \forall k \geq \overline{k}$$

Note that $\frac{1+\alpha}{2+\alpha} \in (\frac{1}{2}, 0)$, as $\alpha = \frac{(C_{KL}\sqrt{2}\mathcal{L}_g)^{\frac{1}{\theta_{KL}}}}{\lambda(1-\theta_{KL})} > 0$. Therefore, we show a linear convergence rate of $\{P_k\}$ i.e. $P_k \leq O((\frac{1+\alpha}{2+\alpha})^{k-\overline{k}}), \ \forall k \geq \overline{k}$. Then using that

$$\|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k - (\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}})\|_F = \|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - \lim_{k\to\infty} (\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\|_F \leq P_k$$

by the triangle inequality, we are guaranteed that the convergence rate is linear:

$$\|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}})\|_F \leq O\left(\left(\frac{1+\alpha}{2+\alpha}\right)^{k-\overline{k}}\right), \quad \forall k \geq \overline{k}$$

Case II: $\theta_{KL} \in (\frac{1}{2}, 1)$**.** In this case, $\frac{1-\theta_{KL}}{\theta_{KL}} \leq 1$. Using a similar analysis as in Case I, we can deduce from eq. (H.7) and the fact $\frac{1-\theta_{KL}}{\theta_{KL}} \leq 1$ to obtain that

$$P_k \leq (1+\alpha)[P_{k-1} - P_k]^{\frac{1-\theta_{KL}}{\theta_{KL}}}, \ \forall k \geq \overline{k}. \tag{H.8}$$

Then, using eq. (H.8) and following a similar argument as in [213, Theorem 2], we get that

$$P_k^{\frac{1-2\theta_{KL}}{1-\theta_{KL}}} - P_{k-1}^{\frac{1-2\theta_{KL}}{1-\theta_{KL}}} \geq \xi, \ \forall k \geq \overline{k} \tag{H.9}$$

for some positive ξ, which implies that

$$P_k^{\frac{1-2\theta_{KL}}{1-\theta_{KL}}} \geq P_k^{\frac{1-2\theta_{KL}}{1-\theta_{KL}}} - P_{\overline{k}-1}^{\frac{1-2\theta_{KL}}{1-\theta_{KL}}} \geq (k-\overline{k})\xi,\ \forall k \geq \overline{k}$$

implying $P_k \leq [(k-\overline{k})\xi]^{-\frac{1-\theta_{KL}}{2\theta_{KL}-1}},\ \forall k \geq \overline{k}$. Finally, using the fact that $\|(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k) - (\overline{\mathbf{U}}, \overline{\mathbf{V}}, \overline{\mathbf{W}})\|_F \leq P_k$, we arrive at a sub-linear convergence rate of $\{(\mathbf{U}_k, \mathbf{V}_k, \mathbf{W}_k)\}$. □