AN ABSTRACT OF THE DISSERTATION OF

Trung Vu for the degree of Doctor of Philosophy in Computer Science presented on September 6, 2022.

Title: Convergence Analysis Framework for Fixed-Point Algorithms in Machine Learning and Signal Processing

Abstract approved: _____

Raviv Raich

Iterative algorithms are simple yet efficient in solving large-scale optimization problems in practice. With a surge in the amount of data in past decades, these methods have become increasingly important in many application areas including matrix/tensor recovery, deep learning, data mining, and reinforcement learning. To optimize or improve iterative algorithms, it is crucial to understand how to characterize their performance. Existing works in the literature offer bounds (including global) on the convergence rate of such algorithms. In most cases, a general global convergence analysis tends to produce conservative convergence rate estimates. In contrast, exact rate analysis predicts accurately the behavior of iterative algorithms in practice. In this dissertation, the goal is to develop a unified framework, with theoretical foundations, to aid the derivation of sharp convergence results for iterative algorithms in machine learning and signal processing (MLSP) problems.

By viewing iterative methods as fixed-point iterations, the existing powerful tools in fixed-point theory are utilized to study their asymptotic convergence. Via the linear approximation of the fixed-point operator around the solution, the proposed approach provides the following key results in convergence analysis: sufficient conditions for local linear convergence, the exact linear rate of convergence, and the number of iterations required to reach certain accuracy. A collection of fundamental MLSP problems are examined to demonstrate the applicability of the proposed framework. In certain problems, such as matrix completion, the novel insight into the local convergence behavior furthers our understanding of the problem and establishes intriguing connections with existing convergence results in the literature. Finally, the dissertation discusses practical methods to obtain the optimal rate of convergence and acceleration techniques that exploit the closed-form expressions of the convergence rate.

Convergence Analysis Framework for Fixed-Point Algorithms in Machine Learning and Signal Processing

by

Trung Vu

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented September 6, 2022
Commencement June 2023

Doctor of Philosophy dissertation of Trung Vu presented on September 6, 2022.

APPROVED:

Major Professor, representing Computer Science

Head of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Trung Vu, Author

ACKNOWLEDGEMENTS

It was around this time six years ago. I remember the younger me with a mixed feeling of excitement and nerves for the first time studying abroad. As the time has flown by, conducting a PhD has been the most challenging journey of my life, with full of ups and downs. I am extremely grateful to the opportunities and experiences it has brought me: to discover and challenge myself, to appraise my strengths and weaknesses, and to build the skills and resilience to become the person I dream of being. I would not have completed this incredible journey without the generous support of many people.

First and foremost, I would like to express my deepest appreciation to my advisor, Dr. Raviv Raich, for his unwavering guidance and patience throughout the entire degree. I very much appreciate the countless hours of discussions with him, from technical details and writing papers to future plans and personal life. Dr. Raich's positive energy and profound belief in my abilities has inspired me with the confidence to complete this dissertation.

I would like to extend my sincere thanks to my committee members, Dr. Xiaoli Fern, Dr. Jinsub Kim, Dr. Thinh Nguyen, and Dr. Matt Campbell, for their unconditional support and valuable advice throughout the duration of my PhD study. Especially helpful to me during this time were Dr. Xiao Fu, Dr. Jinsub Kim, and Dr. Xiaoli Fern, who have contributed their expertise and suggestions to different topics in my research. I also had great pleasure of working with amazing members in the Bioacoustics group, the Smart Vineyards project, and the Sig-

nal Processing group. These include Anh Pham, Evgenia Chunikhina, Zeyu You, Tam Nguyen, Phung Lai, Alan Campbell, Hector Dominguez, Leonardo Machado Cavalcanti, Shashini Akashmika Desilva, Mahtab Aboufazeli, Falah Alanazi, Tri Nguyen, Timothy Marrinan, and many others that I could not list here. I would like to thank all of you and I am hopeful for fruitful collaborations in the future. Additionally, I cannot leave OSU without mentioning the excellent faculty and staff members in the School of EECS. Their enthusiastic support and continued commitment have created a truly inspiring and caring environment for many generations of successful Beavers.

Finally, I am incredibly thankful for my friends and family who have been with me this wonderful journey. I would like to specifically thank the Vietnamese community in Corvallis and Eugene for making my PhD life more colorful and meaningful. I will certainly miss the many heartwarming meals, enjoyable poker nights, and memorable road trips that we had together. My warm thanks should also go to Tam Nguyen, Ngoc Nguyen, Anh Ninh, Jonathan Marcotte, Nana, Bowen, and Xiaohui Lin, who helped me at various stages in my journey. Last but not least, I cannot begin to express my gratitude for my parents, Viet Vu and Ha Nguyen, and my twin brother, Duc Vu. Another six years have already got behind us and without your endless love and constant encouragement, I could not have made it this far. Thank you from the bottom of my heart!

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF ALGORITHMS

Chapter 1: Introduction

1.1 Iterative Algorithms in Machine Learning and Signal Processing

In the era of big data, machine learning and signal processing problems have become increasingly complex. They are often characterized by non-convex geometry, structural constraints, and extremely high dimensions. Representative examples include, but not limited to, sparse recovery [34, 58, 148], matrix completion [33, 42, 44], and phase retrieval [32, 183].

- **Sparse recovery:** Sparse recovery is a classical problem in signal processing in which we wish to acquire and reconstruct a signal efficiently from a series of sampling measurements. In particular, given an n-dimensional signal $\boldsymbol{x}$ that admits sparse/compressible representation either in original domain or in some transform domains (e.g., Fourier transform, cosine transform, wavelet transform), we observe a compressive measurement $\boldsymbol{y}$ of $\boldsymbol{x}$ via an $m \times n$ sensing matrix $\boldsymbol{\Phi}$: $\boldsymbol{y} \approx \boldsymbol{\Phi x}$. Here, the number of measurements taken is much smaller than the length of the input signal, i.e., $m \ll n$. With the introduction of compressed sensing theory [36, 58], the sparsity of $\boldsymbol{x}$ can be exploited to recover $\boldsymbol{x}$ from far fewer samples than required by the Nyquist–Shannon sampling theorem. This is particularly important in systems where measurements are costly such as high-resolution radars [235],

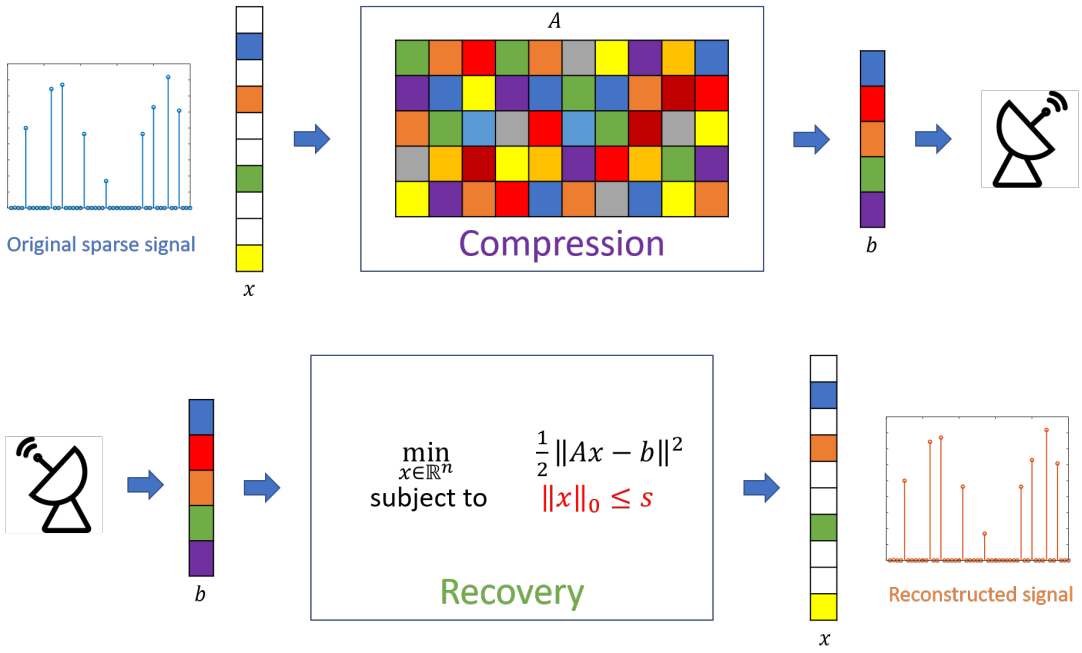


Figure 1.1: A typical setting of sparse recovery. The goal is to recover the sparse signal from a very few number of sampling measurements.

hyper-spectral imaging [98], ECG signal processing [1], and magnetic resonance imaging [163]. Figure 1.1 demonstrates a typical setup of the sparse recovery problem. The sparse recovery problem is often formulated as an L0-norm constrained least squares [20]:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{\Phi x}-\boldsymbol{y}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{x}\|_0 \le s, \tag{1.1}$$

where s is the desired sparsity of the solution. In this formulation, the constraint set is the closed non-convex set of all n-dimensional sparse vectors with at most s non-zero elements, denoted by $\Omega_{\le s}$.

- **Matrix completion:** The matrix completion problem arises in many applications such as collaborative filtering [176, 188, 189, 200], system identification [136, 137, 156], and dimension reduction [31, 228]. Taking a movie recommendation system as an example, we are interested in an $m \times n$ rating matrix $\boldsymbol{M}$ that encodes the preference of m users for n movies (see Fig. 1.2). While such matrix can have thousands to millions of users (rows) and movies (columns), only a handful of the entries of $\boldsymbol{M}$ are available as users typically rate infrequently. To provide good recommendations, it is crucial for the system to make accurate estimates of the unknown entries of $\boldsymbol{M}$ that indicate how each user likes each movie. Additionally, it is reasonable to assume that only a few factors contribute to each user preference (e.g., genre, cast, producer, duration, country, and year). Therefore, the data matrix of all user-ratings may be approximately low-rank. Suppose $\boldsymbol{M}$ is a rank-r that admits a low-rank factorization $\boldsymbol{M} = \boldsymbol{X}^*(\boldsymbol{Y}^*)^\top$, where $\boldsymbol{X}^* \in \mathbb{R}^{m\times r}$ and $\boldsymbol{Y}^* \in \mathbb{R}^{n\times r}$. The problem of recovering the unknown entries of $\boldsymbol{M}$ can be cast as solving a non-convex optimization

$$\min_{\boldsymbol{X}\in\mathbb{R}^{m\times r},\boldsymbol{Y}\in\mathbb{R}^{n\times r}} \frac{1}{2}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}\right)\right\|_F^2, \tag{1.2}$$

where $\mathcal{P}_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is the projection onto the set of matrices sup-

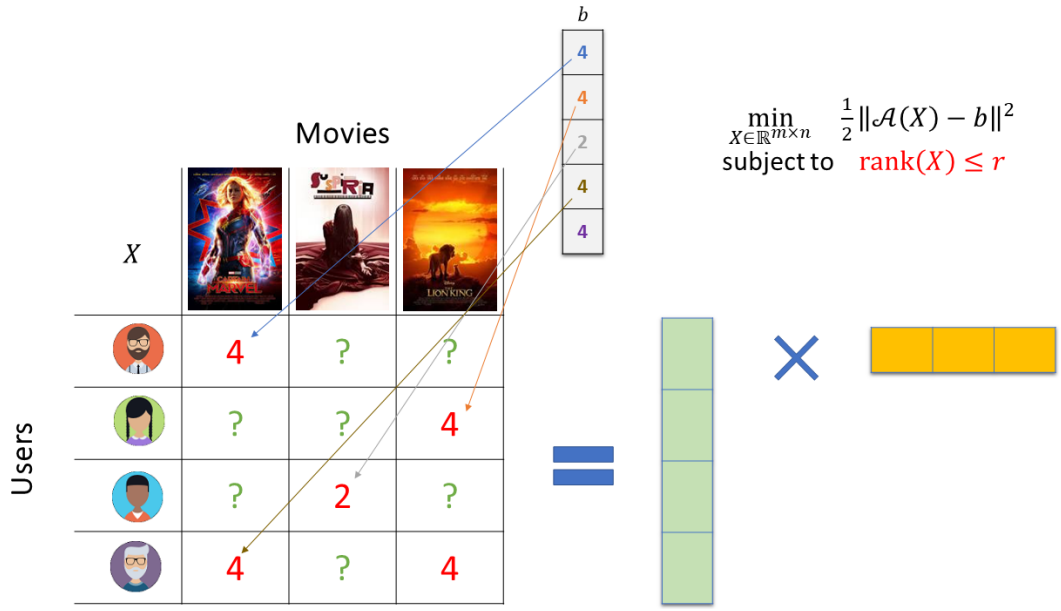


Figure 1.2: A low-rank rating matrix that can be factorized based on latent features from users and movies. The goal of matrix completion is to recover the remaining unobserved ratings in question marks.

ported in Ω, i.e.,

$$[\mathcal{P}_\Omega(\boldsymbol{Z})]_{ij} = \begin{cases} Z_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

- **Phase retrieval:** Phase retrieval is the problem of reconstructing a signal from its Fourier magnitude. This problem arises in many areas of engineering and applied sciences, including X-ray crystallography [152], blind channel estimation [11], optics [222], and speech recognition [170]. In such problems, only measurements of the Fourier magnitude of the underlying signal are

available, while the Fourier phase measurements are missing. Since simply performing an inverse Fourier transform on magnitude measurements without the phase does not recover the original signal successfully (see Fig. 1.3), it is important to come up with algorithms that retrieve the phase from the given magnitude measurements. Formally, phase retrieval can be formulated as[1] the problem of finding a signal $\boldsymbol{x} \in \mathbb{R}^N$ given its Fourier magnitude-square measurements $y_i = \left|\boldsymbol{f}_i^\top \boldsymbol{x}\right|^2$, for $i = 1, \ldots, N$, where $\boldsymbol{f}_i$ is the conjugate of the i-th column of the N-point DFT matrix, with elements $e^{j2\pi in/N}$. In particular, we wish to solve the following least-squares problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} \sum_{i=1}^{N} \left(y_i - \left|\boldsymbol{f}_i^\top \boldsymbol{x}\right|^2\right)^2.$$

In these modern applications, it is crucial to design methods that are numerically efficient, robust against noise, and comes with theoretical guarantees. While second-order methods (e.g., Newton's method) and those dealing with matrix variables (e.g. semidefinite programming) enjoy fast and robust convergence with typically fairly few iterations to reach the desired accuracy, they are computationally prohibitive when the size of the problem increases quickly. On the other hand, lightweight iterative algorithms such as gradient descent and alternating projections have gained a revived interest in large-scale problems thanks to the fact that they are simple to implement as well as require low computational complexity per iteration and small memory storage. Due to the non-convexity in the objective

[1]Here, we focus on the discretized one-dimensional (1D) setting.

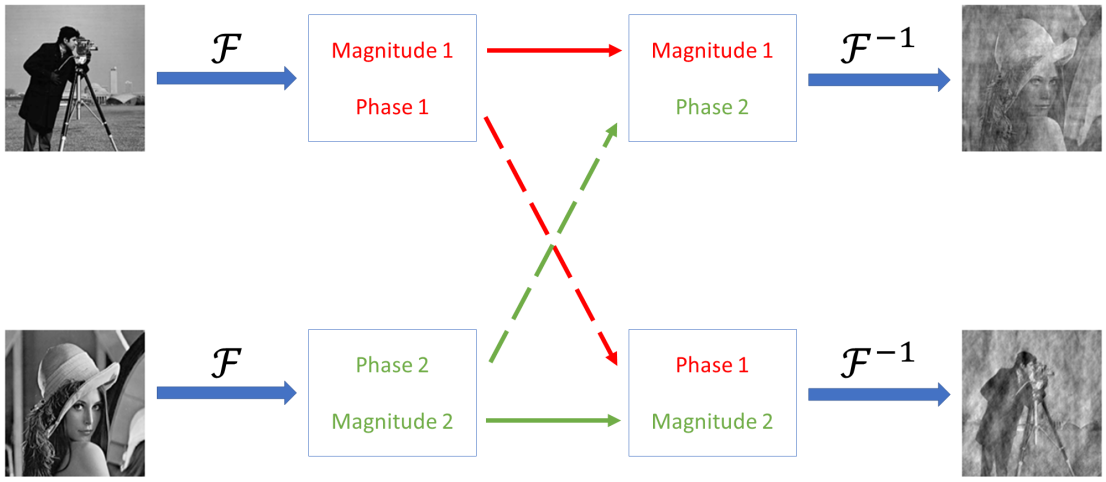


Figure 1.3: An illustration of how the Fourier phase affects image recovery with Fourier transform (reproduction of Fig. 2 in [183]).

functions or the constraints, these problems were initially approached via convex relaxation techniques that are backed by rigorous convergence guarantees for first-order optimization methods [27, 30, 33]. However, researchers soon realized that in practice, the performance of this approach is worse than directly solving the original non-convex problems using first-order methods. Henceforth, there has recently been a shift in focus towards provable and scalable non-convex optimization with representative examples including projected gradient descent, alternating minimizing, and alternating projections.

1.1.1 Iterative Hard Thresholding for Sparse Recovery

The iterative hard thresholding (IHT) algorithm for solving (1.1), which is essentially non-convex projected gradient descent, is based on the following update [20]:

$$\boldsymbol{x}^{(k+1)} = \mathcal{P}_{\Omega_{\leq s}}\big(\boldsymbol{x}^{(k)} - \eta \boldsymbol{\Phi}^{\top}(\boldsymbol{\Phi}\boldsymbol{x}^{(k)} - \boldsymbol{y})\big), \tag{1.3}$$

where $\mathcal{P}_{\Omega_{\leq s}} : \mathbb{R}^n \to \mathbb{R}^n$ is the orthogonal projection onto $\Omega_{\leq s}$. As in [20], $\mathcal{P}_{\Omega_{\leq s}}$ is defined as a non-linear operator that only retains the s coefficients with the largest magnitude:

$$[\mathcal{P}_{\Omega_{\leq s}}(\boldsymbol{z})]_i = \begin{cases} 0 & \text{if } |z_i| < \tau, \\ z_i & \text{if } |z_i| \geq \tau, \end{cases}$$

where τ is set to the smallest magnitude of the s entries in $\boldsymbol{z}$ with largest absolute values. If less than s values are non-zero, we define τ to be the smallest absolute value of the non-zero coefficient.

1.1.2 Factorization-Based Gradient Descent for Matrix Completion

Starting from some guess $(\boldsymbol{X}^{(0)}, \boldsymbol{Y}^{(0)})$, the gradient descent algorithm simply updates the values of $(\boldsymbol{X}, \boldsymbol{Y})$ by taking steps proportional to the negative of the

gradients with respect to each variable:

$$\begin{aligned}\boldsymbol{X}^{(k+1)} &= \boldsymbol{X}^{(k)} - \eta \mathcal{P}_{\Omega}\big(\boldsymbol{X}^{(k)}{\boldsymbol{Y}^{(k)}}^{\top} - \boldsymbol{M}\big)\boldsymbol{Y}^{(k)}, \\ \boldsymbol{Y}^{(k+1)} &= \boldsymbol{Y}^{(k)} - \eta \mathcal{P}_{\Omega}\big(\boldsymbol{X}^{(k)}{\boldsymbol{Y}^{(k)}}^{\top} - \boldsymbol{M}\big)^{\top}\boldsymbol{X}^{(k)},\end{aligned} \tag{1.4}$$

where $\eta > 0$ is the step size.

1.1.3 Alternating Projections for Phase Retrieval

One early approach to phase retrieval is alternating projections, first introduced by Gerchberg and Saxton [78] in 1972. Since the solution must satisfy both the Fourier magnitude constraints and the time domain constraints (e.g., real-valued signals), it can be viewed as the intersection between a convex set (for the time domain constraints) and a non-convex set (for the Fourier magnitude constraints). Thus, the authors proposed to iteratively impose the two set of constraints using projections:

1. Compute the DFT of $\boldsymbol{x}^{(k)}$: $\boldsymbol{z}^{(k+1)} = \boldsymbol{F}\boldsymbol{x}^{(k)}$ and impose the Fourier magnitude constraints $\hat{\boldsymbol{z}}^{(k+1)}(i) = \frac{\boldsymbol{z}^{(k+1)}(i)}{\left|\boldsymbol{z}^{(k+1)}(i)\right|}\sqrt{y_i}$.
2. Compute the inverse DFT of $\hat{\boldsymbol{z}}^{(k+1)}$: $\hat{\boldsymbol{x}}^{(k+1)} = \boldsymbol{F}^{-1}\hat{\boldsymbol{z}}^{(k+1)}$ and impose the time-domain constraints, e.g., $\boldsymbol{x}^{(k+1)} = \mathrm{Re}(\hat{\boldsymbol{x}}^{(k+1)})$.

More concisely, each iteration of this method can be rewritten as

$$\boldsymbol{x}^{(k+1)} = \mathcal{P}_T\big(\boldsymbol{F}^{-1}\mathcal{P}_F(\boldsymbol{F}\boldsymbol{x}^{(k)})\big), \tag{1.5}$$

where $\mathcal{P}_T$ and $\mathcal{P}_F$ are the projections onto the time-domain constraints and the Fourier magnitude constraints, respectively.

1.2 Iterative Algorithms as Fixed-Point Iterations

This dissertation studies the convergence of iterative algorithms as fixed-point iterations, drawing the connection to fixed-point theory for the analysis and design of efficient methods in machine learning and signal processing. More specifically, we represent each update in algorithms like gradient descent and alternating projections as a fixed-point equation of form

$$\boldsymbol{x}^{(k+1)} = \mathcal{F}(\boldsymbol{x}^{(k)}), \tag{1.6}$$

and the convergence of the sequence $\boldsymbol{x}^{(k)}$ can be characterized by the contraction properties and the fixed-points of the operator $\mathcal{F}$.

- **Projected gradient descent:** The IHT update in (1.3) can be viewed as a fixed-point iteration where $\mathcal{F}(\boldsymbol{x}) = \mathcal{P}_{\Omega_{\leq s}}\big(\boldsymbol{x} - \eta\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})\big)$ is a non-smooth non-convex function due to the projection $\mathcal{P}_{\Omega_{\leq s}}$. More generally, we note that many gradient projection based methods can be analyzed under the framework of fixed-point theory, including the singular value projection algorithm for matrix completion [104], the projected gradient descent for unit-modulus least squares in beamforming [206], and the Landweber iteration for solving ill-posed linear inverse problems [48]. In minimizing a differentiable function

f over the constraint set $\mathcal{C}$, the projected gradient descent update is given by

$$\boldsymbol{x}^{(k+1)} = \mathcal{P}_{\mathcal{C}}\big(\boldsymbol{x}^{(k)} - \eta \nabla f(\boldsymbol{x}^{(k)})\big),$$

where $\eta > 0$ is the step size.

- **Gradient descent and its variants:** The gradient descent update in (1.4) can be viewed as a fixed-point iteration where

$$\mathcal{F}(\boldsymbol{Z}) = \begin{bmatrix} \boldsymbol{X} - \eta \mathcal{P}_{\Omega}(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M})\boldsymbol{Y} \\ \boldsymbol{Y} - \eta \mathcal{P}_{\Omega}(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M})^{\top}\boldsymbol{X} \end{bmatrix}, \quad \text{for } \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix},$$

  is a smooth non-convex function with respect to $\boldsymbol{Z}$. This view also applied to other contexts such as the Wirtinger flow algorithm for phase retrieval [32]. In convex optimization, the fixed-point view of gradient descent and its variants has been studied by Jung [113].

- **Alternating projections:** The alternating projection update in (1.5) can be viewed as a fixed-point iteration where $\mathcal{F}(\boldsymbol{x}) = \mathcal{P}_T\big(\boldsymbol{F}^{-1}\mathcal{P}_F(\boldsymbol{F}\boldsymbol{x})\big)$ is a smooth non-convex function due to the projection $\mathcal{P}_F$. Similarly, the same perspective can be applied to alternating projections of form

$$\boldsymbol{x}^{(k+1)} = \mathcal{P}_{\mathcal{C}_1}\big(\ldots \mathcal{P}_{\mathcal{C}_m}(\boldsymbol{x}^{(k)})\big),$$

  where $\mathcal{C}_1, \ldots, \mathcal{C}_m$ are the constraint sets and $\mathcal{P}_{\mathcal{C}_1}, \ldots, \mathcal{P}_{\mathcal{C}_m}$ are the correspond-

ing projections onto them. Such update arises naturally in other problems such as matrix completion [46], color plane interpolation [84], and source localization [238].

By the fixed-point theorem [121], if the Jacobian of $\mathcal{F}$ is bounded uniformly, in some natural matrix norm, by $\rho \in (0,1)$, the sequence $\boldsymbol{x}^{(k)}$ generated by (1.6) converges linearly to a fixed-point $\boldsymbol{x}^*$ of $\mathcal{F}$ at rate $\rho_k \leq \rho$:

$$\left\|\boldsymbol{x}^{(l+1)} - \boldsymbol{x}^*\right\| \leq \rho_k \left\|\boldsymbol{x}^{(l)} - \boldsymbol{x}^*\right\| \qquad \text{for } l = k, k+1, \ldots$$

Furthermore, as the iterates converge, ρ_k approaches $\rho(J_{\mathcal{F}}(\boldsymbol{x}^*))$, the spectral radius of the Jacobian matrix at the fixed point. It is interesting here to emphasize that such fixed-point results are powerful tools to study the convergence of iterative algorithms in MLSP, especially the asymptotic convergence with exact linear rate. However, little research for convergence of iterative algorithms in the MLSP literature has been done in this direction. To further motivate our interpretation of iterative algorithms as fixed-point iterations, let us briefly review the existing convergence results for iterative algorithms.

1.3 Asymptotic Convergence of Iterative Algorithms

From a theoretical point of view, convergence properties of iterative algorithms have long been studied. These properties involve two key aspects: the quality of convergent points and the speed of convergence. On the one hand, the quality

of convergent points provides useful insights into when the algorithm converges, whether it converges to a global/local optimum or a stationary/critical point, and how (far) the objective function at the convergent point compares to the optimal objective value. On the other hand, the speed of convergence concerns the order of convergence, the rate of convergence, and the number of iterations required to obtain sufficiently small errors. In this dissertation, the focus is on the second aspect that measures the efficiency of iterative algorithms, in particular, the asymptotic rate of convergence. To better understand the concept of convergence rate, consider gradient descent as one representative of iterative algorithms. In order to minimize a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the algorithm, starting from some initial guess $\boldsymbol{x}^{(0)}$, performs the following iterative update

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \eta \nabla f(\boldsymbol{x}^{(k)}), \tag{1.7}$$

where $\eta > 0$ is the step size (a.k.a, the learning rate). Assume $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ converges to $\boldsymbol{x}^*$. Then, the convergence of $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ to $\boldsymbol{x}^*$ is said to be at rate μ if there exists a bounding sequence $\{\epsilon_k\}_{k=0}^{\infty}$ such that $\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|_2 \leq \epsilon_k$ for all k and $\lim_{k\to\infty} \epsilon_{k+1}/\epsilon_k = \mu$. The asymptotic rate of convergence of gradient descent to $\boldsymbol{x}^*$, denoted by ρ, is defined by the worst-case rate of convergence among all possible sequences $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ that are generated by (1.7) and converge to $\boldsymbol{x}^*$, i.e., $\rho = \sup_{\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}} \mu$. Depending on the value of ρ in the interval $[0, 1]$, the convergence is said to be sublinear ($\rho = 1$), linear ($0 < \rho < 1$) or superlinear ($\rho = 0$). It is straightforward that the lower the value of ρ is, the faster the speed of convergence

is; and typically fewer iterations are necessary to obtain a close approximation of the solution. Thus, analytical estimation of the convergence rate plays a pivotal role in convergence analysis.

In analyzing the convergence of iterative methods, it is common to study the convergence to a global solution of the problem. Via assumptions on the strong convexity and the smoothness of the problem, this approach primarily focuses on the quality of convergent points (global versus local) and the order of convergence. It provides a universal upper bound on the error reduction at each k-th iteration, which holds for both the asymptotic ($k \to \infty$) and non-asymptotic (small k) convergence regime [12, 16, 24, 103, 160]. The disadvantage of this analysis, however, is that it often underestimate the asymptotic rate due to the conservative nature of the employed bounding techniques. A lesser-known approach to convergence rate analysis is to establish the exact asymptotic rate of convergence by exploiting the local structure of the problem. By focusing on the local behavior of iterative algorithms near the solution, this approach offers sharper results on the convergence rate, particularly in the case of non-quadratic objectives. Dating back to the 1960s, there are two major methods for asymptotic convergence rate analysis. The first method was proposed by Polyak in [167], based on his earlier study into nonlinear difference equations [166]. The key ideas in this approach are the extension of the mean value theorem to vector-valued functions and the stability of difference equations. The second method for asymptotic convergence rate analysis was developed by Daniel [51] in 1967, while studying gradient descent with exact line search, i.e., choosing η that minimizes the objective at each iteration.

Utilizing the Kantorovich inequality [114], the author proved a similar result on convergence characteristics that are close to those inherent for quadratics are exploited through the Hessian $\nabla^2 f(\boldsymbol{x}^*)$. The same technique was then extended to study the asymptotic convergence of projected gradient descent for constrained optimization [71, 132, 139].

Apart from the asymptotic convergence rate, one would also be interested in the region of convergence and the number of iterations needed to reach certain accuracy. Both of the existing methods, nonetheless, do not provide further result on these aspects of convergence. Moreover, to the best of our knowledge, there has been no extension of Polyak's method to the case of projected gradient descent (with fixed step size scheme), and vice versa, there has been no extension of Daniel's method beyond the exact line search scheme.

1.4 Focus Areas

1.4.1 Asymptotic Convergence Analysis

Motivated by the fixed-point view of iterative algorithms, this dissertation aims to develop a unified framework to study their asymptotic convergence, namely, the convergence rate, the region of convergence, and the number of required iterations. Differently from Polyak's approach to the stability of non-linear difference equation in [166], we approximate the fixed-point function (locally) by the following first-

order difference equation

$$\boldsymbol{\delta}^{(k+1)} = T(\boldsymbol{\delta}^{(k)}) + \boldsymbol{q}(\boldsymbol{\delta}^{(k)}), \tag{1.8}$$

where $\boldsymbol{\delta}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^*$ is the residual at the k-th iteration, T is a linear operator that acts as a contraction mapping on $\boldsymbol{\delta}$, and $\boldsymbol{q} : \mathbb{R}^n \to \mathbb{R}^n$ satisfies $\lim_{\|\boldsymbol{\delta}\|\to 0} \|\boldsymbol{q}(\boldsymbol{\delta})\|_2 / \|\boldsymbol{\delta}\|_2^2 < \infty$. By carefully examining the stability of the system dynamic (1.8), the fundamental requirements to achieve the local linear convergence rate can be identified. The methodology applies to both gradient descent and projected gradient descent with fixed step sizes, as well as iterative algorithms whose updates can be represented as fixed-point iterations, e.g., alternating projections. This author hopes that the proposed framework will be used by researchers in the area of MLSP as a general recipe to quickly derive sharp convergence results for their specific problems. A collection of fundamental statistical estimation problems will also be examined to demonstrate the applicability of the proposed framework. In some of the applications, the novel insight into the local convergence reveals interesting connections to existing results in relevant areas such as global convergence analysis, random matrix theory, perturbation theory, and differential geometry.

1.4.2 Acceleration Techniques

Another contribution of this dissertation is that the insight into asymptotic convergence analysis can be used to develop variants that enjoy faster convergence while remain the same computational complexity per iteration. The simplest approach is to select the optimal step size based on the closed-form expression of the local convergence rate obtained by the proposed framework. Such selection can be used as a benchmark against practical schemes with adaptive step size like backtracking line search. In a more elaborated approach, acceleration techniques such as the Heavy-Ball method and Nesterov's accelerated gradient have been introduced in optimization literature as well as have been using widely in practice. However, with the fixed-point view of iterative algorithms, these techniques arise naturally via exploiting leveraging the well-known results in fixed-point theory [181, 221]. Moreover, based on the convergence analysis of the plain algorithm, one can easily design the accelerated variant with optimal parameters (e.g., momentum step size).

The chapters of this dissertation follow our published work or work under review. For the ease of the readers, the chapters are self-contained, following closely the corresponding publication/manuscript. An overview of the dissertation is shown in Fig. 1.4. The rest of this document is organized as follows. Chapter 2 introduces a closed-form bound on the convergence of iterative methods via fixed point analysis. This serves as a mathematical tool for the subsequence analysis of convergence, establishing the expressions of the convergence rate, region of conver-

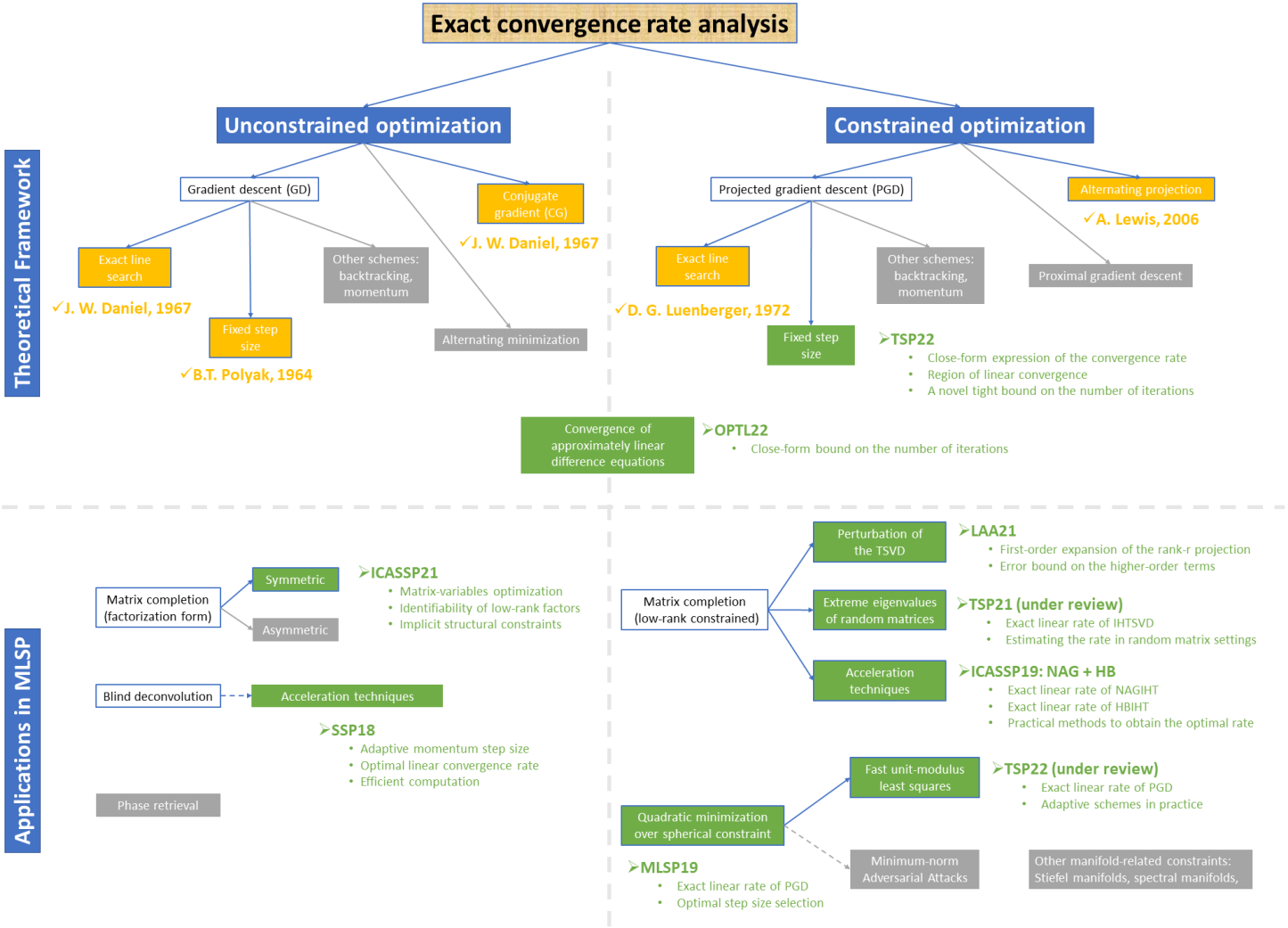


Figure 1.4: An overview of this dissertation. The research focuses on the convergence rate analysis for iterative algorithms and its application in machine learning and signal processing. The yellow boxes indicate existing works in the literature, the green boxes indicate works that were completed in this dissertation, and the gray boxes indicate future works.

gence, and the number of required iterations to reach certain accuracy. Chapter 3 presents a unified framework to study the local linear convergence of projected gradient descent in the general context of constrained least squares. Then, the application of the proposed framework is demonstrated for the following problems: minimizing a quadratic over a sphere (Chapter 4), unit-modulus constrained least squares (Chapter 5), and low-rank matrix completion (Chapter 7). Focusing on the latter problem, we present a handful of analytical results on the rank-r projection operator (Chapter 6), the extreme eigenvalues of random matrices, and their connections to the asymptotic convergence of iterative hard thresholding for matrix completion. We also demonstrate some acceleration techniques for IHT that can used to exploit the asymptotic convergence results and obtain the optimal convergence in practice (Chapters 8 and 9). Another method for matrix completion, gradient descent for the factorization-based formulation, is also analyzed in Chapter 10 under the view of fixed-point iterations. Chapter 11 concludes the technical part of the dissertation with the study of an adaptive step size schedule for momentum methods in deconvolution applications. Finally, Chapter 12 summarizes the contribution of the dissertation and discusses potential directions for future work.

Chapter 2: A Closed-Form Bound on the Asymptotic Linear Convergence of Iterative Methods via Fixed Point Analysis[1]

In many iterative optimization methods, fixed-point theory enables the analysis of the convergence rate via the contraction factor associated with the linear approximation of the fixed-point operator. While this factor characterizes the asymptotic linear rate of convergence, it does not explain the non-linear behavior of these algorithms in the non-asymptotic regime. In this chapter, we take into account the effect of the first-order approximation error and present a closed-form bound on the convergence in terms of the number of iterations required for the distance between the iterate and the limit point to reach an arbitrarily small fraction of the initial distance. Our bound includes two terms: one corresponds to the number of iterations required for the linearized version of the fixed-point operator and the other corresponds to the overhead associated with the approximation error. With a focus on the convergence in the scalar case, the tightness of the proposed bound is proven for positively quadratic first-order difference equations.

[1]This work has been published as: Trung Vu and Raviv Raich. "A Closed-Form Bound on the Asymptotic Linear Convergence of Iterative Methods via Fixed Point Analysis." Optimization Letters, vol. 1, pp. 1-14, 2022.

2.1 Introduction

Many iterative optimization methods, such as gradient descent and alternating projections, can be interpreted as fixed-point iterations [113, 166, 181, 221]. Such methods consist of the construction of a series $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty} \subset \mathbb{R}^n$ generated by

$$\boldsymbol{x}^{(k+1)} = \mathcal{F}(\boldsymbol{x}^{(k)}), \tag{2.1}$$

where the fixed-point operator $\mathcal{F}$ is an endomorphism on $\mathbb{R}^n$. By the fixed-point theorem [9, 25, 121], if the Jacobian of $\mathcal{F}$ is bounded uniformly, in the matrix norm $\|\cdot\|_2$ induced by the Euclidean norm for vectors $\|\cdot\|$, by $\rho \in (0, 1)$, the sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ generated by (2.1) converges locally to a fixed-point $\boldsymbol{x}^*$ of $\mathcal{F}$ at a linear rate ρ, i.e., $\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\right\| \leq \rho \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|$ for all integer k.[2] Assume that $\mathcal{F}$ is differentiable at $\boldsymbol{x}^*$ and admits the first-order expansion [178]

$$\mathcal{F}(\boldsymbol{x}^{(k)}) = \mathcal{F}(\boldsymbol{x}^*) + \mathcal{T}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^*) + \boldsymbol{q}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^*),$$

where $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ is the derivative of $\mathcal{F}$ at $\boldsymbol{x}^*$ and $\boldsymbol{q} : \mathbb{R}^n \to \mathbb{R}^n$ is the residual satisfying $\limsup_{\boldsymbol{\delta} \to 0} \|\boldsymbol{q}(\boldsymbol{\delta})\| / \|\boldsymbol{\delta}\| = 0$. Then, denoting the error at the k-th iteration as $\boldsymbol{\delta}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^*$, the fixed-point iteration (2.1) can be viewed as a non-linear but approximately linear difference equation

$$\boldsymbol{\delta}^{(k+1)} = \mathcal{T}(\boldsymbol{\delta}^{(k)}) + \boldsymbol{q}(\boldsymbol{\delta}^{(k)}). \tag{2.2}$$

[2] $\|\cdot\|$ denotes the Euclidean norm.

The stability of non-linear difference equations of form (2.2) has been studied by Polyak [166] in 1964, extending the result from the continuous domain [14]. In particular, the author showed that if the spectral radius of $\mathcal{T}$, denoted by $\rho(\mathcal{T})$, is strictly less than 1, then for arbitrarily small $\zeta > 0$, there exists a constant $C(\zeta)$ such that $\left\|\boldsymbol{\delta}^{(k)}\right\| \leq C(\zeta)\left\|\boldsymbol{\delta}^{(0)}\right\|(\rho(\mathcal{T})+\zeta)^k$ with sufficiently small $\left\|\boldsymbol{\delta}^{(0)}\right\|$. While this result characterizes the asymptotic linear convergence of (2.2), it does not specify the exact conditions on how small $\left\|\boldsymbol{\delta}^{(0)}\right\|$ is as well as how large the factor $C(\zeta)$ is.

This chapter develops a more elaborate approach to analyze the convergence of (2.2) that offers, in addition to the asymptotic linear rate $\rho(\mathcal{T})$, both the region of convergence (i.e., a set $\mathcal{S}$ such that for any $\boldsymbol{\delta}^{(0)} \in \mathcal{S}$ we have $\lim_{k\to\infty}\left\|\boldsymbol{\delta}^{(k)}\right\| = 0$) and a tight closed-form bound on $H(\epsilon)$ defined as the smallest integer guaranteeing $\left\|\boldsymbol{\delta}^{(k)}\right\| \leq \epsilon\left\|\boldsymbol{\delta}^{(0)}\right\|$ for $0 < \epsilon < 1$ and all $k \geq H(\epsilon)$. We begin with the scalar version of (2.2) in which the residual term $q(\delta)$ is replaced with an exact quadratic function of δ and then extend the result to the original vector case. In the first step, we study the convergence of the sequence $\{a_k\}_{k=0}^{\infty} \subset \mathbb{R}$, generated by the following quadratic first-order difference equation

$$a_{k+1} = \rho a_k + q a_k^2, \tag{2.3}$$

where $a_0 > 0$, $0 < \rho < 1$, and $q \geq 0$ are real scalars. In the second step, we consider the sequence $\{a_k\}_{k=0}^{\infty}$ obtained by (2.3) with $\rho = \rho(\mathcal{T})$ and $q = \sup_{\boldsymbol{\delta}\in\mathbb{R}^n} \frac{\|\boldsymbol{q}(\boldsymbol{\delta})\|}{\|\boldsymbol{\delta}\|^2}$ as an upper bound for the sequence $\{\left\|\boldsymbol{\delta}^{(k)}\right\|\}_{k=0}^{\infty}$. In this chapter, we focus on the former

step while the latter step is obtained using a more straightforward derivation.

In analyzing the convergence of $\{a_k\}_{k=0}^{\infty}$, we focus on tightly characterizing $K(\epsilon)$ (for $0 < \epsilon < 1$), which is defined as the smallest integer such that $a_k \leq \epsilon a_0$ for all $k \geq K(\epsilon)$. The value of $K(\epsilon)$ serves as an upper bound on $H(\epsilon)$. When $q = 0$, (2.3) becomes a linear first-order difference equation and $\{a_k\}_{k=0}^{\infty}$ converges uniformly to 0 at a **linear rate** ρ. In particular, $a_{k+1} = \rho a_k$ implies $a_k = a_0 \rho^k$ for any non-negative integer k. Then, for $q = 0$, an exact expression of $K(\epsilon)$ can be obtained in closed-form as

$$K(\epsilon) = \left\lceil \frac{\log(1/\epsilon)}{\log(1/\rho)} \right\rceil. \tag{2.4}$$

When $q > 0$, the sequence $\{a_k\}_{k=0}^{\infty}$ either converges, diverges or remains constant depending on the initial value a_0:

1. If $a_0 > (1-\rho)/q$, then $\{a_k\}_{k=0}^{\infty}$ diverges.

2. If $a_0 = (1-\rho)/q$, then $a_k = (1-\rho)/q$ for all $k \in \mathbb{N}$.

3. If $a_0 < (1-\rho)/q$, then $\{a_k\}_{k=0}^{\infty}$ converges to 0 monotonically.

We are interested in the convergence of the sequence $\{a_k\}_{k=0}^{\infty}$ for $a_0 < (1-\rho)/q$. In the asymptotic regime (a_k small), the convergence is almost linear since the first-order term ρa_k dominates the second-order term $q a_k^2$. In the early stage (a_k large), on the other hand, the convergence is non-linear due to the strong effect of $q a_k^2$. In addition, when $\rho \to 0$, one would expect $\{a_k\}_{k=0}^{\infty}$ enjoys a fast quadratic convergence as $q a_k^2$ dominates ρa_k. On the other end of the spectrum, when $\rho \to$

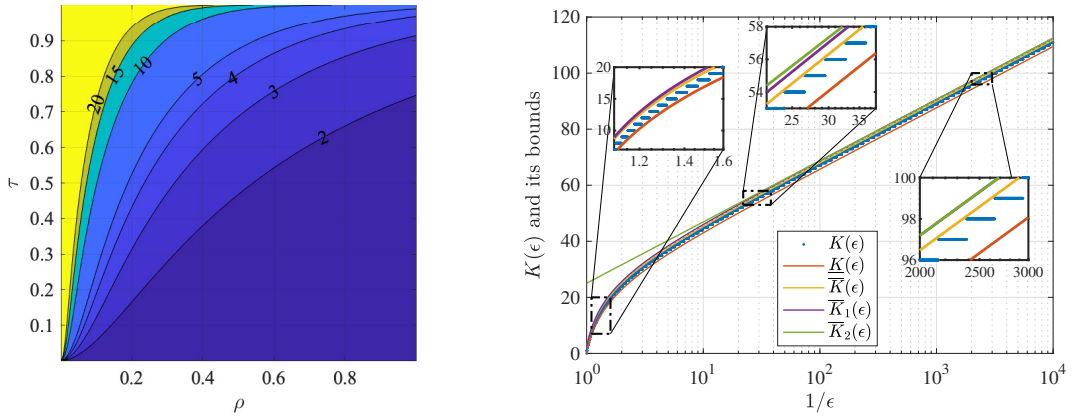


Figure 2.1: (Left) Contour plot of the bound on asymptotic gap between $\overline{K}_2(\epsilon)$ and $K(\epsilon)$, given in (2.8). (Right) Log-scale plot of $K(\epsilon)$ and its bounds as functions of $1/\epsilon$, with $\rho = 0.9$ and $\tau = 0.89$. Three zoomed plots are added to the original plot for better visualization.

1, we observe that the convergence is even slower than linear, making it more challenging to estimate $K(\epsilon)$.

2.2 Asymptotic Convergence in the Scalar Case

In this section, we provide a tight upper bound on $K(\epsilon)$ in terms of a_0, ρ, q, and ϵ. Our bound suggests the sequence $\{a_k\}_{k=0}^{\infty}$ converges to 0 at an **asymptotically linear rate** ρ with an overhead cost that depends on only two quantities: ρ and $a_0 q/(1-\rho)$. Our main result is stated as follows.

Theorem 2.1. *Consider the sequence $\{a_k\}_{k=0}^{\infty}$ defined in (2.3) with $a_0 > 0$, $0 < \rho < 1$, and $q > 0$. Assume that $a_0 < (1-\rho)/q$ and denote $\tau = a_0 q/(1-\rho)$ (where $0 < \tau < 1$). Then, for any $0 < \epsilon < 1$, the smallest integer, denoted by $K(\epsilon)$, such*

that $a_k \leq \epsilon a_0$ for all $k \geq K(\epsilon)$, can be bounded as follows

$$K(\epsilon) \leq \frac{\log(1/\epsilon)}{\log(1/\rho)} + c(\rho,\tau) \triangleq \overline{K}_2(\epsilon), \tag{2.5}$$

where

$$c(\rho,\tau) = \frac{1}{\rho\log(1/\rho)}\Delta E_1\Big(\log\frac{1}{\rho+\tau(1-\rho)}, \log\frac{1}{\rho}\Big) + b(\rho,\tau), \tag{2.6}$$

$\Delta E_1(x,y) = E_1(x) - E_1(y)$, $E_1(x) = \int_x^\infty \frac{e^{-t}}{t}dt$ is the exponential integral [2], and

$$b(\rho,\tau) = \frac{1}{2\rho}\log\bigg(\frac{\log(1/\rho)}{\log\big(1/(\rho+\tau(1-\rho))\big)}\bigg) + 1. \tag{2.7}$$

Moreover, the gap $\overline{K}_2(\epsilon) - K(\epsilon)$ is upper-bounded asymptotically as follows[3]

$$\lim_{\epsilon\to 0}\Big(\overline{K}_2(\epsilon) - K(\epsilon)\Big) \leq \frac{\Delta E_1\big(2\log\frac{1}{\rho+\tau(1-\rho)}, 2\log\frac{1}{\rho}\big) - \rho\Delta E_1\big(\log\frac{1}{\rho+\tau(1-\rho)}, \log\frac{1}{\rho}\big)}{2\rho^2\log(1/\rho)} + b(\rho,\tau). \tag{2.8}$$

The proof of Theorem 2.1 is given in Appendix 2.5. The upper bound $\overline{K}_2(\epsilon)$, given in (2.5), is the sum of two terms: (i) the first term is similar to (2.4), representing the asymptotic linear convergence of $\{a_k\}_{k=0}^\infty$; (ii) the second term, $c(\rho,\tau)$, is independent of ϵ, representing the overhead in the number of iterations caused by the non-linear term qa_k^2. This overhead term is understood as the additional number of iterations beyond the number of iterations for the linear model. As one would

[3]A tighter version of the upper bound $\overline{K}_2(\epsilon)$ is given in Appendix 2.5, cf., (2.19) and (2.21).

expect, when $a_0 \to (1-\rho)/q$, we have $\tau \to 1$ and $c(\rho,\tau)$ approaches infinity. On the other hand, when $\tau \to 0$, the gap from the number of iterations required by the linear model $c(\rho,\tau)$ approaches 1. The right hand side (RHS) of (2.8) is an upper bound on the asymptotic gap between our proposed upper bound on $K(\epsilon)$ and the actual value of $K(\epsilon)$ and hence represents the tightness of our bound. The value of the bound as a function of ρ and τ is shown in Fig. 2.1 (left). It is notable that the asymptotic gap is guaranteed to be no more than 10 iterations for a large portion of the (ρ,τ)-space. It is particularly small in the lower right part of the figure. For example, for $\rho \geq 0.9$ and $\tau \leq 0.9$, the gap is no more than 4 iterations. Figure 2.1 (right) demonstrates different bounds on $K(\epsilon)$ (blue dotted line) including $\overline{K}_2(\epsilon)$ (green solid line). We refer the readers to Appendix 2.5 for the details of other bounds in the figure. We observe that the upper bound $\overline{K}_2(\epsilon)$ approaches $K(\epsilon)$ as $\epsilon \to 0$, with the asymptotic gap of less than 2 iterations. On the other hand, $\overline{K}_2(\epsilon)$ reaches $c(\rho,\tau) \approx 25$ as $\epsilon \to 1$, suggesting that the proposed bound $\overline{K}_2(\epsilon)$ requires no more than 25 iterations beyond the number of iterations required by the linear model to achieve $a_k \leq \epsilon a_0$.

2.3 Extension to the Vector Case

We now consider an extension of Theorem 2.1 to the convergence analysis in the vector case given by (2.2). More elaborate applications of the proposed analysis in convergence analysis of iterative optimization methods can be found in [213–215, 218].

Theorem 2.2. *Consider the difference equation*

$$\boldsymbol{\delta}^{(k+1)} = \boldsymbol{T}\boldsymbol{\delta}^{(k)} + \boldsymbol{q}(\boldsymbol{\delta}^{(k)}), \tag{2.9}$$

where $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ *admits an eigendecomposition* $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$, $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ *is an invertible matrix with the condition number* $\kappa(\boldsymbol{Q}) = \|\boldsymbol{Q}\|_2 \|\boldsymbol{Q}^{-1}\|_2$, *and* $\boldsymbol{\Lambda}$ *is an* $n \times n$ *diagonal matrix whose entries are strictly less than* 1 *in magnitude. In addition, assume that there exists a finite constant* $q > 0$ *satisfying* $\|\boldsymbol{q}(\boldsymbol{\delta})\| \le q\|\boldsymbol{\delta}\|^2$ *for any* $\boldsymbol{\delta} \in \mathbb{R}^n$. *Then, for any* $0 < \epsilon < 1$, *we have* $\|\boldsymbol{\delta}^{(k)}\| \le \epsilon \|\boldsymbol{\delta}^{(0)}\|$ *provided that*

$$\|\boldsymbol{\delta}^{(0)}\| < \frac{1-\rho(\boldsymbol{T})}{q\kappa(\boldsymbol{Q})^2} \text{ and } k \ge \frac{\log(1/\epsilon) + \log(\kappa(\boldsymbol{Q}))}{\log(1/\rho(\boldsymbol{T}))} + c\big(\rho(\boldsymbol{T}), \frac{q\kappa(\boldsymbol{Q})\|\boldsymbol{Q}\|_2 \|\boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(0)}\|}{1-\rho(\boldsymbol{T})}\big) \tag{2.10}$$

where $c(\rho, \tau)$ *is given in (2.6). Moreover, if* $\boldsymbol{T}$ *is symmetric, then (2.10) becomes*

$$\|\boldsymbol{\delta}^{(0)}\| < \frac{1-\rho(\boldsymbol{T})}{q} \text{ and } k \ge \frac{\log(1/\epsilon)}{\log(1/\rho(\boldsymbol{T}))} + c\Big(\rho(\boldsymbol{T}), \frac{q\|\boldsymbol{\delta}^{(0)}\|}{1-\rho(\boldsymbol{T})}\Big). \tag{2.11}$$

Note that the RHS of the inequalities involving k in both (2.10) and (2.11) serve as upper bounds to $H(\epsilon)$ defined in the introduction. Moreover, the sets of all $\boldsymbol{\delta}^{(0)}$ that satisfy the inequality involving $\boldsymbol{\delta}^{(0)}$ in both (2.10) and (2.11) offer valid regions of convergence. Similar to the scalar case, we observe in the number of required iterations one term corresponding to the asymptotic linear convergence and another term corresponding to the non-linear convergence at the early stage. When $\boldsymbol{T}$ is asymmetric, there is an additional cost of diagonalizing $\boldsymbol{T}$, associated

with $\kappa(\boldsymbol{Q})$ in (2.10). The proof of Theorem 2.2 is given in Appendix 2.6.

2.4 Conclusion

With a focus on fixed-point iterations, we analyzed the convergence of the sequence generated by a quadratic first-order difference equation. We presented a bound on the minimum number of iterations required for the distance between the iterate and the limit point to reach an arbitrarily small fraction of the initial distance. Our bound includes two terms: one corresponds to the number of iterations required for the linearized difference equation and the other corresponds to the overhead associated with the residual term. The bound for the vector case is derived based on a tight bound obtained for the scalar quadratic difference equation. A characterization of the tightness of the bound for the scalar quadratic difference equation was introduced.

2.5 Proof of Theorem 2.1

First, we establish a sandwich inequality on $K(\epsilon)$ in the following lemma:

Lemma 2.1. *For any $0 < \epsilon < 1$, let $K(\epsilon)$ be the smallest integer such that for all $k \geq K(\epsilon)$, we have $a_k \leq \epsilon a_0$. Then,*

$$\underline{K}(\epsilon) \triangleq F\big(\log(1/\epsilon)\big) \leq K(\epsilon) \leq F\big(\log(1/\epsilon)\big) + b(\rho, \tau) \triangleq \overline{K}(\epsilon), \tag{2.12}$$

where $b(\rho,\tau)$ is defined in (2.7) and

$$F(x)=\int_0^x f(t)dt \quad \textit{with} \quad f(x)=\frac{1}{-\log\big(\rho+\tau(1-\rho)e^{-x}\big)}. \tag{2.13}$$

The lemma provides an upper bound on $K(\epsilon)$. Moreover, it is a tight bound in the sense that the gap between lower bound $\underline{K}(\epsilon)$ and the upper bound $\overline{K}(\epsilon)$ is independent of ϵ. In other words, the ratio $K(\epsilon)/\overline{K}(\epsilon)$ approaches 1 as $\epsilon\to 0$. Next, we proceed to obtain a tight closed-form upper bound on $\overline{K}(\epsilon)$ by upper-bounding $F(\log(1/\epsilon))$.

Lemma 2.2. *Consider the function $F(\cdot)$ given in (2.13). For $0<\epsilon<1$, we have*

$$F\big(\log(1/\epsilon)\big)\le \frac{\log(1/\epsilon)}{\log(1/\rho)}+\frac{\Delta E_1\Big(\log\frac{1}{\rho+\tau(1-\rho)},\log\frac{1}{\rho+\epsilon\tau(1-\rho)}\Big)}{\rho\log(1/\rho)}\triangleq \overline{F}_1\big(\log(1/\epsilon)\big) \tag{2.14}$$

$$\le \frac{\log(1/\epsilon)}{\log(1/\rho)}+\frac{\Delta E_1\Big(\log\frac{1}{\rho+\tau(1-\rho)},\log\frac{1}{\rho}\Big)}{\rho\log(1/\rho)}\triangleq \overline{F}_2\big(\log(1/\epsilon)\big) \tag{2.15}$$

and

$$F\big(\log(1/\epsilon)\big)\ge \overline{F}_1\big(\log(1/\epsilon)\big)-A(\epsilon)\triangleq \underline{F}_1\big(\log(1/\epsilon)\big), \tag{2.16}$$

where

$$A(\epsilon) \triangleq \frac{\Delta E_1\big(2\log\frac{1}{\rho+\tau(1-\rho)}, 2\log\frac{1}{\rho+\tau(1-\rho)\epsilon}\big) - \rho\Delta E_1\big(\log\frac{1}{\rho+\tau(1-\rho)}, \log\frac{1}{\rho+\tau(1-\rho)\epsilon}\big)}{2\rho^2\log(1/\rho)}. \tag{2.17}$$

Lemma 2.2 offers two upper bounds on $F(\log(1/\epsilon))$ and one lower bound. The first bound $\overline{F}_1(\log(1/\epsilon))$ approximates well the behavior of $F(\log(1/\epsilon))$ for both small and large values of $\log(1/\epsilon)$. The second bound $\overline{F}_2(\log(1/\epsilon))$ provides a linear bound on $F(\log(1/\epsilon))$ in terms of $\log(1/\epsilon)$. Moreover, the gap between $F(\log(1/\epsilon))$ and $\underline{F}_1(\log(1/\epsilon))$, given by $A(\epsilon)$, can be upper bound by $A(0)$ since $A(\cdot)$ is monotonically decreasing for $\epsilon \in [0, 1)$. While $F(\cdot)$ asymptotically increases like $\log(1/\epsilon)/\log(1/\rho)$, the gap approaches a constant independent of ϵ. Replacing $F(\log(1/\epsilon))$ on the RHS of (2.12) by either of the upper bounds in Lemma 2.2, we obtain two corresponding bounds on $K(\epsilon)$:

$$\overline{K}_1(\epsilon) \triangleq \overline{F}_1\big(\log(1/\epsilon)\big) + b(\rho,\tau) \leq \overline{F}_2\big(\log(1/\epsilon)\big) + b(\rho,\tau) \triangleq \overline{K}_2(\epsilon), \tag{2.18}$$

where we note that $\overline{K}_2(\epsilon)$ has the same expression as in (2.5). Moreover, the tightness of these two upper bounds can be shown as follows. First, using the first inequality in (2.12) and then the lower bound on $F(\log(1/\epsilon))$ in (2.16), the gap

between $\overline{K}_1(\epsilon)$ and $K(\epsilon)$ can be bounded by

$$\begin{aligned}
\overline{K}_1(\epsilon) - K(\epsilon) &\leq \overline{K}_1(\epsilon) - F\big(\log(1/\epsilon)\big) \\
&\leq \overline{K}_1(\epsilon) - \Big(\overline{F}_1\big(\log(1/\epsilon)\big) - A(\epsilon)\Big) \\
&= \Big(\overline{F}_1\big(\big(\log(1/\epsilon)\big) + b(\rho,\tau)\Big) - \Big(\overline{F}_1\big(\log(1/\epsilon)\big) - A(\epsilon)\Big) \\
&= A(\epsilon) + b(\rho,\tau) \\
&\leq A(0) + b(\rho,\tau), \qquad (2.19)
\end{aligned}$$

where the last inequality stems from the monotonicity of $A(\cdot)$ in $[0,1)$. Note that the bound in (2.19) holds uniformly independent of ϵ, implying $\overline{K}_1(\epsilon)$ is a tight bound on $K(\epsilon)$. Second, using (2.18), the gap between $\overline{K}_2(\epsilon)$ and $K(\epsilon)$ can be represented as

$$\begin{aligned}
\overline{K}_2(\epsilon) - K(\epsilon) &= \big(\overline{K}_2(\epsilon) - \overline{K}_1(\epsilon)\big) + \big(\overline{K}_1(\epsilon) - K(\epsilon)\big) \\
&= \big(\overline{F}_2\big(\log(1/\epsilon)\big) - \overline{F}_1\big(\log(1/\epsilon)\big)\big) + \big(\overline{K}_1(\epsilon) - K(\epsilon)\big) \\
&\leq \big(\overline{F}_2\big(\log(1/\epsilon)\big) - \overline{F}_1\big(\log(1/\epsilon)\big)\big) + \big(A(0) + b(\rho,\tau)\big), \qquad (2.20)
\end{aligned}$$

where the last inequality stems from (2.19). Furthermore, using the definition of $\overline{F}_1(\log(1/\epsilon))$ and $\overline{F}_2(\log(1/\epsilon))$ in (2.14) and (2.15), respectively, we have

$$\lim_{\epsilon \to 0}(\overline{F}_2(\log(1/\epsilon)) - \overline{F}_1(\log(1/\epsilon))) = 0.$$

Thus, taking the limit $\epsilon \to 0$ on both sides of (2.20), we obtain

$$\lim_{\epsilon \to 0} \big(\overline{K}_2(\epsilon) - K(\epsilon) \big) \leq A(0) + b(\rho, \tau). \tag{2.21}$$

We note that $\overline{K}_2(\epsilon)$ is a simple bound that is linear in terms of $\log(1/\epsilon)$ and approaches the upper bound $\overline{K}_1(\epsilon)$ in the asymptotic regime ($\epsilon \to 0$). Evaluating $A(0)$ from (2.17) and substituting it back into (2.21) yields (2.8), which completes our proof of Theorem 2.1. Figure 2.1 (right) depicts the aforementioned bounds on $K(\epsilon)$. It can be seen from the plot that all the four bounds match the asymptotic rate of increment in $K(\epsilon)$ (for large values of $1/\epsilon$). The three bounds $\underline{K}(\epsilon)$ (red), $\overline{K}(\epsilon)$ (yellow), and $\overline{K}_1(\epsilon)$ (purple) closely follow $K(\epsilon)$ (blue), indicating that the integral function $F(\cdot)$ effectively estimates the minimum number of iterations required to achieve $a_k \leq \epsilon a_0$ in this setting. The upper bound $\overline{K}_2(\epsilon)$ (green) forms a tangent to $\overline{K}_1(\epsilon)$ at $1/\epsilon \to \infty$ (i.e., $\epsilon \to 0$).

2.5.1 Proof of Lemma 2.1

Let $d_k = \log(a_0/a_k)$ for each $k \in \mathbb{N}$. Substituting $a_k = a_0 e^{-d_k}$ into (2.3), we obtain the surrogate sequence $\{d_k\}_{k=0}^{\infty}$:

$$d_{k+1} = d_k - \log\big(\rho + \tau(1-\rho)e^{-d_k}\big), \tag{2.22}$$

where $d_0 = 0$ and $\tau = a_0 q/(1-\rho) \in (0,1)$. Since $\{a_k\}_{k=0}^{\infty}$ is monotonically decreasing to 0 and d_k is monotonically decreasing as a function of a_k, $\{d_k\}_{k=0}^{\infty}$ is a

monotonically increasing sequence. Our key steps in this proof are first to tightly bound the index $K \in \mathbb{N}$ using $F(d_K)$

$$F(d_K) \leq K \leq F(d_K) + \frac{1}{2\rho} \log\left(\frac{\log \rho}{\log\big(\rho + \tau(1-\rho)\big)} \right) \tag{2.23}$$

and then to obtain (2.12) from (2.23) using the monotonicity of the sequence $\{d_k\}_{k=0}^{\infty}$ and of the function $F(\cdot)$. We proceed with the details of each of the steps in the following.

Step 1: We prove (2.23) by showing the lower bound on K first and then showing the upper bound on K. Using (2.13), we can rewrite (2.22) as $d_{k+1} = d_k + 1/f(d_k)$. Rearranging this equation yields

$$f(d_k)(d_{k+1} - d_k) = 1. \tag{2.24}$$

Since $f(x)$ is monotonically decreasing, we obtain the lower bound on K in (2.23) by

$$\begin{aligned} F(d_K) &= \int_0^{d_K} f(x)dx = \sum_{k=0}^{K-1} \int_{d_k}^{d_{k+1}} f(x)dx \\ &\leq \sum_{k=0}^{K-1} \int_{d_k}^{d_{k+1}} f(d_k)dx = \sum_{k=0}^{K-1} f(d_k)(d_{k+1} - d_k) = K, \end{aligned} \tag{2.25}$$

where the last equality stems from (2.24). For the upper bound on K in (2.23),

we use the convexity of $f(\cdot)$ to lower-bound $F(d_K)$ as follows

$$\begin{aligned} F(d_K) &= \sum_{k=0}^{K-1} \int_{d_k}^{d_{k+1}} f(x)dx \geq \sum_{k=0}^{K-1} \int_{d_k}^{d_{k+1}} \big(f(d_k) + f'(d_k)(x-d_k)\big)dx \\ &= \sum_{k=0}^{K-1} \Big(f(d_k)(d_{k+1}-d_k) + \frac{1}{2}f'(d_k)(d_{k+1}-d_k)^2\Big). \end{aligned} \tag{2.26}$$

Using (2.24) and substituting $f'(x) = -\big(f(x)\big)^2 \frac{\tau(1-\rho)e^{-x}}{\rho+\tau(1-\rho)e^{-x}}$ into the RHS of (2.26), we obtain

$$F(d_K) \geq K - \frac{1}{2}\sum_{k=0}^{K-1} \frac{\tau(1-\rho)e^{-d_k}}{\rho+\tau(1-\rho)e^{-d_k}}. \tag{2.27}$$

Note that (2.27) already offers an upper on K in terms of $F(d_K)$. To obtain the upper bound on K in (2.23) from (2.27), it suffices to show that

$$\sum_{k=0}^{K-1} \frac{\tau(1-\rho)e^{-d_k}}{\rho+\tau(1-\rho)e^{-d_k}} \leq \frac{1}{\rho}\log\Big(\frac{\log\rho}{\log\big(\rho+\tau(1-\rho)\big)}\Big). \tag{2.28}$$

In the following, we prove (2.28) by introducing the functions

$$g(x) = \frac{\tau(1-\rho)e^{-x}}{\rho+\tau(1-\rho)e^{-x}} \frac{1}{-\log\big(\rho+\tau(1-\rho)e^{-x}\big)} \tag{2.29}$$

and

$$G(x) = \int_0^x g(t)dt = \log\Big(\frac{\log(\rho+\tau(1-\rho)e^{-x})}{\log(\rho+\tau(1-\rho))}\Big). \tag{2.30}$$

Note that $g(\cdot)$ is monotonically decreasing (a product of two decreasing functions) while $G(\cdot)$ is monotonically increasing (an integral of a non-negative function) on $[0, \infty)$. We have

$$\begin{aligned} G(d_K) &= \int_0^{d_K} g(x)dx = \sum_{k=0}^{K-1} \int_{d_k}^{d_{k+1}} g(x)dx \geq \sum_{k=0}^{K-1} \int_{d_k}^{d_{k+1}} g(d_{k+1})dx \\ &= \sum_{k=0}^{K-1} g(d_{k+1})(d_{k+1} - d_k) = \sum_{k=0}^{K-1} \frac{g(d_{k+1})}{g(d_k)} g(d_k)(d_{k+1} - d_k). \end{aligned} \tag{2.31}$$

Lemma 2.3. *For any $k \in \mathbb{N}$, we have $g(d_{k+1})/g(d_k) \geq \rho$.*

Proof. For $k \in \mathbb{N}$, let $t_k = \rho + \tau(1-\rho)e^{-d_k} \in (\rho, 1)$. From (2.22), we have $t_k = e^{-(d_{k+1}-d_k)}$ and $t_{k+1} = \rho + \tau(1-\rho)e^{-d_{k+1}} = \rho + \tau(1-\rho)e^{-d_k}e^{-(d_{k+1}-d_k)} = \rho + (t_k - \rho)t_k$. Substituting d_k for x in $g(x)$ from (2.29) and replacing $\rho + \tau(1-\rho)e^{-d_k}$ with t_k yield $g(d_k) = \frac{\tau(1-\rho)e^{-d_k}}{t_k} \frac{1}{-\log(t_k)}$. Repeating the same process to obtain $g(d_{k+1})$ and taking the ratio between $g(d_{k+1})$ and $g(d_k)$, we obtain

$$\frac{g(d_{k+1})}{g(d_k)} = e^{-(d_{k+1}-d_k)} \frac{t_k}{t_{k+1}} \frac{\log(t_k)}{\log(t_{k+1})}. \tag{2.32}$$

Substituting $e^{-(d_{k+1}-d_k)} = t_k$ and $t_{k+1} = \rho + (t_k - \rho)t_k$ into (2.32) yields

$$\frac{g(d_{k+1})}{g(d_k)} = \frac{t_k^2 \log(t_k)}{(\rho + (t_k - \rho)t_k) \log(\rho + (t_k - \rho)t_k)}. \tag{2.33}$$

We now continue to bound the ratio $g(d_{k+1})/g(d_k)$ by bounding the RHS. Since $t_k - \rho \geq 0$ and $t_k < 1$, we have $t_k - \rho > (t_k - \rho)t_k$ and hence $t_k/(\rho + (t_k - \rho)t_k) > 1$. Thus, in order to prove $\frac{g(d_{k+1})}{g(d_k)} \geq \rho$ from the fact that the RHS of (2.33) is greater

or equal to ρ, it remains to show that

$$\frac{t_k \log(t_k)}{\log\big(\rho + (t_k - \rho)t_k\big)} \geq \rho. \tag{2.34}$$

By the concavity of $\log(\cdot)$, it holds that $\log(\frac{\rho}{t_k}1 + \frac{t_k-\rho}{t_k}t) \geq \frac{\rho}{t_k}\log(1) + \frac{t_k-\rho}{t_k}\log(t_k) = (1 - \frac{\rho}{t_k})\log(t_k)$. Adding $\log(t_k)$ to both sides of the last inequality yields $\log(\rho + (t_k - \rho)t_k) \geq (2 - \frac{\rho}{t_k})\log(t_k)$. Now using the fact that $(\sqrt{\rho/t_k} - \sqrt{t_k/\rho})^2 \geq 0$, we have $2 - \rho/t_k \leq t_k/\rho$. By this inequality and the negativity of $\log(t_k)$, we have $\log(\rho + (t_k - \rho)t_k) \geq \frac{t_k}{\rho}\log(t_k)$. Multiplying both sides by the negative ratio $\rho/\log(\rho + (t_k - \rho)t_k)$ and adjusting the direction of the inequality yields the inequality in (2.34), which completes our proof of the lemma. □

Back to our proof of Theorem 2.1, applying Lemma 2.3 to (2.31) and substituting $d_{k+1} - d_k = -\log(\rho + \tau(1-\rho)e^{-d_k})$ from (2.22) and $g(d_k)$ from (2.29), we have

$$G(d_K) \geq \sum_{k=0}^{K-1} \rho g(d_k)(d_{k+1} - d_k) = \rho \sum_{k=0}^{K-1} \frac{\tau(1-\rho)e^{-d_k}}{\rho + \tau(1-\rho)e^{-d_k}}. \tag{2.35}$$

Using the monotonicity of $G(\cdot)$, we upper-bound $G(d_K)$ by

$$G(d_K) \leq G(\infty) = \log\Big(\frac{\log \rho}{\log(\rho + \tau(1-\rho))}\Big). \tag{2.36}$$

Thus, the RHS of (2.35) is upper bounded by the RHS of (2.36). Dividing the result by ρ, we obtain (2.28). This completes our proof of the upper bound on K in (2.23) and thereby the first step of the proof.

Step 2: We proved both the lower bound and the upper bound on K in (2.23). Next, we proceed to show (2.12) using (2.23). By the definition of $K(\epsilon)$, $a_{K(\epsilon)} \leq \epsilon a_0 < a_{K(\epsilon)-1}$. Since $d_k = \log(a_0/a_k)$, for $k \in \mathbb{N}$, we have $d_{K(\epsilon)-1} \leq \log(1/\epsilon) \leq d_{K(\epsilon)}$. On the one hand, using the monotonicity of $F(\cdot)$ and substituting $K = K(\epsilon)$ into the lower bound on K in (2.23) yields

$$F\big(\log(1/\epsilon)\big) \leq F(d_{K(\epsilon)}) \leq K(\epsilon). \tag{2.37}$$

On the other hand, substituting $K = K(\epsilon) - 1$ into the upper bound on K in (2.23), we obtain

$$K(\epsilon) - 1 \leq F(d_{K(\epsilon)-1}) + \frac{1}{2\rho} \log\left(\frac{\log \rho}{\log\big(\rho + \tau(1-\rho)\big)}\right). \tag{2.38}$$

Since $F(\cdot)$ is monotonically increasing and $d_{K(\epsilon)-1} \leq \log(1/\epsilon)$, we have $F(d_{K(\epsilon)-1}) \leq F(\log(1/\epsilon))$. Therefore, upper-bounding $F(d_{K(\epsilon)-1})$ on the RHS of (2.38) by $F(\log(1/\epsilon))$ yields

$$K(\epsilon) \leq F\big(\log(1/\epsilon)\big) + \frac{1}{2\rho} \log\left(\frac{\log \rho}{\log\big(\rho + \tau(1-\rho)\big)}\right) + 1. \tag{2.39}$$

The inequality (2.12) follows on combining (2.37) and (2.39).

2.5.2 Proof of Lemma 2.2

Let $\nu = \tau(1-\rho)/\rho$. We represent $f(x)$ in the interval $(0, \log(1/\epsilon))$ as

$$\begin{aligned} f(x) &= \frac{1}{-\log\big(\rho + \tau(1-\rho)e^{-x}\big)} \\ &= \frac{1}{\log(1/\rho)} + \frac{1}{\log(1/\rho)} \frac{\log(1+\nu e^{-x})}{\log(1/\rho) - \log(1+\nu e^{-x})}. \end{aligned}$$

Then, taking the integral from 0 to $\log(1/\epsilon)$ yields

$$F\big(\log(1/\epsilon)\big) = \frac{1}{\log(1/\rho)} \left(\log(1/\epsilon) + \int_0^{\log(1/\epsilon)} \frac{\log(1+\nu e^{-t})}{\log(1/\rho) - \log(1+\nu e^{-t})} dt \right). \tag{2.40}$$

Using $\alpha(1-\alpha/2) = \alpha - \alpha^2/2 \le \log(1+\alpha) \le \alpha$, for $\alpha = \nu e^{-t} \ge 0$, on the numerator within the integral in (2.40) and changing the integration variable t to $z = \log(1/\rho) - \log(1+\nu e^{-t})$, we obtain both an upper bound and a lower bound on the integral on the RHS of (2.40)

$$\begin{aligned} \frac{1}{\rho} \int_{\underline{z}}^{\overline{z}} \frac{e^{-z} - \frac{1}{2\rho} e^{-z}(e^{-z} - \rho)}{z} dz &\le \int_0^{\log(1/\epsilon)} \frac{\log(1+\nu e^{-t})}{\log(1/\rho) - \log(1+\nu e^{-t})} dt \\ &\le \frac{1}{\rho} \int_{\underline{z}}^{\overline{z}} \frac{e^{-z}}{z} dz, \end{aligned} \tag{2.41}$$

where $\underline{z} = -\log(\rho + \tau(1-\rho))$ and $\overline{z} = -\log(\rho + \epsilon\tau(1-\rho))$. Replacing the integral in (2.40) by the upper bound and lower bound from (2.41), using the definition of the exponential integral, and simplifying, we obtain the upper-bound on $F(\log(1/\epsilon))$ given by $\overline{F}_1(\log(1/\epsilon))$ in (2.14) and similarly the lower bound on $F(\log(1/\epsilon))$

given by $\underline{F}_1(\log(1/\epsilon))$ in (2.16). Finally, we prove the second upper bound in (2.15) as follows. Since $E_1(\cdot)$ is monotonically decreasing and $\frac{1}{\rho+\epsilon\tau(1-\rho)} \leq \frac{1}{\rho}$, we have $E_1(\log \frac{1}{\rho+\epsilon\tau(1-\rho)}) \geq E_1(\log \frac{1}{\rho})$, which implies $\Delta E_1(\log \frac{1}{\rho+\tau(1-\rho)}, \log \frac{1}{\rho+\epsilon\tau(1-\rho)}) \leq \Delta E_1(\log \frac{1}{\rho+\tau(1-\rho)}, \log \frac{1}{\rho})$. Combining this with the definition of $\overline{F}_1(\log(1/\epsilon))$ and $\overline{F}_2(\log(1/\epsilon))$ in (2.14) and (2.15), respectively, we conclude that $\overline{F}_1(\log(1/\epsilon)) \leq \overline{F}_2(\log(1/\epsilon))$, thereby completes the proof of the lemma.

2.6 Proof of Theorem 2.2

Let $\tilde{\boldsymbol{\delta}}^{(k)} = \boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(k)}$ be the transformed error vector. Substituting $\mathcal{T}(\boldsymbol{\delta}^{(k)}) = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}(\boldsymbol{\delta}^{(k)})$ into (2.2) and then left-multiplying both sides by $\boldsymbol{Q}^{-1}$, we obtain

$$\tilde{\boldsymbol{\delta}}^{(k+1)} = \boldsymbol{\Lambda}\tilde{\boldsymbol{\delta}}^{(k)} + \tilde{\boldsymbol{q}}(\tilde{\boldsymbol{\delta}}^{(k)}), \tag{2.42}$$

where $\tilde{\boldsymbol{q}}(\tilde{\boldsymbol{\delta}}^{(k)}) = \boldsymbol{Q}^{-1}\boldsymbol{q}(\boldsymbol{Q}\tilde{\boldsymbol{\delta}}^{(k)})$ satisfies $\left\|\tilde{\boldsymbol{q}}(\tilde{\boldsymbol{\delta}}^{(k)})\right\| \leq q \left\|\boldsymbol{Q}^{-1}\right\|_2 \left\|\boldsymbol{Q}\right\|_2^2 \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|^2$. Taking the norm of both sides of (2.42) and using the triangle inequality yield

$$\begin{aligned}\left\|\tilde{\boldsymbol{\delta}}^{(k+1)}\right\| &\leq \left\|\boldsymbol{\Lambda}\tilde{\boldsymbol{\delta}}^{(k)}\right\| + \left\|\tilde{\boldsymbol{q}}(\tilde{\boldsymbol{\delta}}^{(k)})\right\| \\ &\leq \left\|\boldsymbol{\Lambda}\right\|_2 \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| + q \left\|\boldsymbol{Q}^{-1}\right\|_2 \left\|\boldsymbol{Q}\right\|_2^2 \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|^2\end{aligned}$$

Since $\left\|\boldsymbol{\Lambda}\right\|_2 = \rho(\mathcal{T})$, the last inequality can be rewritten compactly as

$$\left\|\tilde{\boldsymbol{\delta}}^{(k+1)}\right\| \leq \rho \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| + \tilde{q} \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|^2, \tag{2.43}$$

where $\rho = \rho(\mathcal{T})$ and $\tilde{q} = q \left\|\boldsymbol{Q}^{-1}\right\|_2 \left\|\boldsymbol{Q}\right\|_2^2$.

To analyze the convergence of $\{\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|\}_{k=0}^{\infty}$, let us consider a surrogate sequence $\{a_k\}_{k=0}^{\infty} \subset \mathbb{R}$ defined by $a_{k+1} = \rho a_k + \tilde{q} a_k^2$ with $a_0 = \left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\|$. We show that $\{a_k\}_{k=0}^{\infty}$ upper-bounds $\{\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|\}_{k=0}^{\infty}$, i.e.,

$$\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| \leq a_k \quad \forall k \in \mathbb{N}. \tag{2.44}$$

The base case when $k = 0$ holds trivially as $a_0 = \left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\|$. In the induction step, given $\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| \leq a_k$ for some integer $k \geq 0$, we have

$$\left\|\tilde{\boldsymbol{\delta}}^{(k+1)}\right\| \leq \rho \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| + \tilde{q} \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|^2 \leq \rho a_k + \tilde{q} a_k^2 = a_{k+1}.$$

By the principle of induction, (2.44) holds for all $k \in \mathbb{N}$. Assume for now that $a_0 = \left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\| < (1-\rho)/\tilde{q}$, then applying Theorem 2.1 yields $a_k \leq \tilde{\epsilon} a_0$ for any $\tilde{\epsilon} > 0$ and integer $k \geq \log(1/\tilde{\epsilon})/\log(1/\rho) + c(\rho, \tau)$. Using (2.44) and setting $\tilde{\epsilon} = \epsilon/\kappa(\boldsymbol{Q})$, we further have $\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| \leq a_k \leq \tilde{\epsilon} a_0 = \epsilon \left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\| / \kappa(\boldsymbol{Q})$ for all

$$k \geq \frac{\log(1/\epsilon) + \log(\kappa(\boldsymbol{Q}))}{\log(1/\rho)} + c\Big(\rho, \frac{\tilde{q} \left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\|}{1-\rho}\Big). \tag{2.45}$$

Now, it remains to prove *(i)* the accuracy on the transformed error vector $\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| \leq \tilde{\epsilon} \left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\|$ is sufficient for the accuracy on the original error vector $\left\|\boldsymbol{\delta}^{(k)}\right\| \leq \epsilon \left\|\boldsymbol{\delta}^{(0)}\right\|$; and *(ii)* the initial condition $\left\|\boldsymbol{\delta}^{(0)}\right\| < (1-\rho)/(q\kappa(\boldsymbol{Q})^2)$ is sufficient for $\left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\| <$

$(1-\rho)/\tilde{q}$. In order to prove *(i)*, using $\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| \leq \epsilon \left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\| / \kappa(\boldsymbol{Q})$, we have

$$\begin{aligned}\left\|\boldsymbol{\delta}^{(k)}\right\| = \left\|\boldsymbol{Q}\tilde{\boldsymbol{\delta}}^{(k)}\right\| &\leq \|\boldsymbol{Q}\|_2 \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| \\ &\leq \|\boldsymbol{Q}\|_2 \frac{\epsilon}{\|\boldsymbol{Q}\|_2 \|\boldsymbol{Q}^{-1}\|_2} \left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\| = \frac{\epsilon}{\|\boldsymbol{Q}^{-1}\|_2} \left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\| \leq \epsilon \left\|\boldsymbol{\delta}^{(0)}\right\|,\end{aligned}$$

where the last inequality stems from $\left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\| = \left\|\boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(0)}\right\| \leq \|\boldsymbol{Q}^{-1}\|_2 \left\|\boldsymbol{\delta}^{(0)}\right\|$. To prove *(ii)*, we use similar derivation as follows

$$\left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\| \leq \left\|\boldsymbol{Q}^{-1}\right\|_2 \left\|\boldsymbol{\delta}^{(0)}\right\| < \left\|\boldsymbol{Q}^{-1}\right\|_2 \frac{1-\rho}{q\kappa(\boldsymbol{Q})^2} = \frac{1-\rho}{\tilde{q}}.$$

Finally, the case that $\boldsymbol{T}$ is symmetric can be proven by the fact that $\boldsymbol{Q}$ is orthogonal, i.e., $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T$ and $\kappa(\boldsymbol{Q}) = 1$. Substituting this back into (2.10) and using the orthogonal invariance property of norm, we obtain the simplified version in (2.11). This completes our proof of Theorem 2.2.

Chapter 3: On Local Linear Convergence of Projected Gradient Descent for Constrained Least Squares[1]

Many recent problems in signal processing and machine learning such as compressed sensing, image restoration, matrix/tensor recovery, and non-negative matrix factorization can be cast as constrained optimization. Projected gradient descent is a simple yet efficient method for solving such constrained optimization problems. Local convergence analysis furthers our understanding of its asymptotic behavior near the solution, offering sharper bounds on the convergence rate compared to global convergence analysis. However, local guarantees often appear scattered in problem-specific areas of machine learning and signal processing. This chapter presents a unified framework for the local convergence analysis of projected gradient descent in the context of constrained least squares. The proposed analysis offers insights into pivotal local convergence properties such as the conditions for linear convergence, the region of convergence, the exact asymptotic rate of convergence, and the bound on the number of iterations needed to reach a certain level of accuracy. To demonstrate the applicability of the proposed approach, we present a recipe for the convergence analysis of projected gradient descent and demonstrate

[1]This work has been published as: Trung Vu and Raviv Raich. "On Local Linear Convergence of Projected Gradient Descent for Constrained Least Squares." IEEE Transactions on Signal Processing, vol. 70, pp. 4061-4076, 2022.

it via a beginning-to-end application of the recipe on four fundamental problems, namely, linear equality-constrained least squares, sparse recovery, least squares with the unit norm constraint, and matrix completion.

3.1 Introduction

Constrained least squares can be formulated as the following optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2 \quad \text{s.t. } \boldsymbol{x}\in\mathcal{C}, \tag{3.1}$$

where $\mathcal{C}\in\mathbb{R}^n$ is a non-empty closed set, $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, and $\boldsymbol{b}\in\mathbb{R}^m$ is the observation from which we wish to recover the solution $\boldsymbol{x}^*$ efficiently. With the surge in the amount of data over the past decades, modern learning problems have become increasingly complex and optimization in the presence of constraints is frequently used to capture accurately their inherent structure. Examples in the area of machine learning and signal processing include, but are not limited to, compressed sensing [20, 21, 66], image restoration [72, 100, 150], seismic inversion [41, 149, 169], and phase-only beamforming [206, 236]. Since the set of real $n_1\times n_2$ matrices is isomorphic to $\mathbb{R}^{n_1 n_2}$, application of (3.1) is also found in problems such as low-rank matrix recovery [43, 104, 116] and non-negative matrix factorization [83, 133, 154].

Projected gradient descent (PGD) is one of the most popular methods for solving constrained optimization, thanks to its simplicity and efficiency. In theory, convergence properties of this method are natural extensions of the classical re-

sults for unconstrained optimization [12, 16, 103, 140]. When the constraint set $\mathcal{C}$ is convex, PGD is also known as the projected Landweber iteration [48] and is shown to converge sublinearly to the global solution of (3.1). Moreover, when the least-squares objective is strongly convex, the algorithm enjoys fast linear convergence. For non-convex settings, with the recent introduction of restricted (strong) convexity, global convergence has been guaranteed for certain structural constraints such as sparsity constraint [36], low-rank constraint [199], and L2-norm constraint [13]. For a more comprehensive review of convergence analysis for PGD in the literature, we refer the reader to Appendix 3.6.6.

From a different perspective, problem (3.1) can be viewed as a manifold optimization problem in which the intrinsic structure of manifolds can be exploited. Dating back to the 1970s, Luenberger [139] studied a variant of gradient projection method using the concept of geodesic descent. Under the assumption that $\mathcal{C}$ is a differentiable manifold in Euclidean space, the author provided sufficient conditions for global convergence and established a sharp bound on the asymptotic convergence rate near a strict local minimum. Later on, this result was extended to a broader class of Riemannian manifolds and has been widely known as the Riemannian steepest descent method [71, 132, 139, 207]. The asymptotic convergence rate of Riemannian steepest descent (with exact line search) is given by the Kantorovich ratio $(\beta - \alpha)^2/(\beta + \alpha)^2$, where α and β are the smallest and largest eigenvalues of the second derivative of the Lagrangian restricted to the subspace tangent to the constraint manifold at the solution. Remarkably, such local convergence bounds are tighter than those obtained from the aforementioned global

convergence analysis in the optimization literature since the former exploits the local structure of the problem. The global convergence bounds, on the other hand, take into account the worst-case behavior of the algorithm that might occur far away from the solution of interest. In certain situations, global convergence analysis suggests sublinear convergence while local convergence analysis offers linear convergence thanks to the benign structure near the solution [164]. One key element in the asymptotic convergence analysis of Riemannian steepest descent is Kantorovich inequality [197]. However, this technique depends on the optimal choice of step size in the exact line search scheme and is not straightforwardly generalized to other variants of gradient projection. To the best of our knowledge, there has been no direct extension of the analysis for Riemannian steepest descent method to plain PGD with a fixed step size.

Our Contribution. In this chapter, we develop a unified framework for a local convergence analysis of the PGD algorithm. We leverage our earlier preliminary work, in which we developed a convergence rate only analysis for the specific problems of low-rank matrix completion [46] and minimization of a quadratic with spherical constraints [218]. For the former, we developed two acceleration approaches that leverage on the rate analysis [213, 214]. The key approach used in these works is to represent each algorithm as a fixed point iteration and to approximate the fixed point operator as locally linear. This idea extends to other algorithms (i.e., non PGD) that can be represented using a fixed point iteration (e.g., see our work on analyzing GD for symmetric matrix completion [215]). For each problem, problem-specific properties have been utilized to facilitate the anal-

ysis. Here, our goal is to develop a *unified* framework for convergence rate analysis of PGD for constrained least-squares. Our framework relies on three key steps: *(i)* the introduction of Lipschitz-continuous differentiability to provide tight error bounds on the linear approximation of the projection operator near the solution, *(ii)* the establishment of an asymptotically-linear recursion on the error iterations, and *(iii)* the derivation of the linear rate and the region of convergence (ROC) of the error sequence by leveraging our work on the convergence of nonlinear difference equations [216]. Our approach shifts the burden of the analysis to the characterization of the projection operator (for an example of such characterization of the projection onto the rank-r manifold, see [211]-Theorem 1). In the context of PGD for the general constrained least squares, the proposed framework is the first to offer a closed-form expression of the exact asymptotic rate of local linear convergence, the ROC, and a bound on the number of iterations needed to reach a certain level of accuracy.[2] To illustrate the utility of the approach, we apply our framework to four well-known problems in machine learning and signal processing, namely, linear equality-constrained least squares, sparse recovery, least squares with spherical constraint, and matrix completion. We show that the obtained asymptotic rate of convergence matches existing results in the literature. For problems in which the exact convergence rate of PGD has not been studied, we

[2]We note that the classic work of Polyak [166] can be considered as a replacement for our analysis in the third step. While such result is more general in the context of nonlinear different equations, we do not find a straightforward extension to obtain the ROC and the guarantees on the number of required iterations in our context of convergence analysis.

verify the asymptotic rate obtained by our analysis against the rate of convergence obtained in numerical experiments. We believe that this framework can be used as a general recipe to develop quick yet sharp local convergence results for PGD in other applications in the field as well as to complement conservative analysis of global convergence.

Organization. The rest of this chapter is organized as follows. Section 3.2 provides a brief background of PGD for constrained least squares, including properties of the orthogonal projection, stationary points of the problem, and the PGD algorithm along with its fixed points. Next, we present our unified framework for the local convergence analysis of PGD in Section 3.3, followed by the proof of the main theorem. Then, Section 3.4 demonstrates the application of the proposed recipe to four well-known problems in machine learning and signal processing. Finally, we summarize our results and discuss some of the possible extensions in Section 3.5.

3.2 Preliminaries

This section presents key concepts and background results that will be used as the basic premise of our subsequent convergence analysis.

3.2.1 Notation

Throughout this chapter, we use the notation $\|\cdot\|$ to denote the Euclidean norm for vectors. For matrices, $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the Frobenius norm and the spectral norm, respectively. Boldfaced symbols are reserved for vectors and matrices. Additionally, the $t \times t$ identity matrix is denoted by $\boldsymbol{I}_t$ and the ith vector in the natural basis of $\mathbb{R}^n$ is denoted by $\boldsymbol{e}_i$. We use $\otimes$ to denote the Kronecker product between two matrices. The vectorization of a matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, denoted by $\mathrm{vec}(\boldsymbol{X})$, is the concatenation of the columns of a matrix one on top of another in their original order, i.e., for $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, $\mathrm{vec}(\boldsymbol{X}) = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top]^\top$. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$, $\mathrm{diag}(\boldsymbol{x})$ denotes the a square diagonal matrix such that $[\mathrm{diag}(\boldsymbol{x})]_{ii} = x_i$. For a scalar $r > 0$, denote the open ball of center $\boldsymbol{x}$ and radius r by $\mathcal{B}(\boldsymbol{x}, r) = \{\boldsymbol{y} \mid \|\boldsymbol{y} - \boldsymbol{x}\| < r\}$. Correspondingly, the closed ball of center $\boldsymbol{x}$ and radius r is denoted by $\mathcal{B}[\boldsymbol{x}, r] = \{\boldsymbol{y} \mid \|\boldsymbol{y} - \boldsymbol{x}\| \le r\}$. The lexicographical order between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same length is defined by $\boldsymbol{x} < \boldsymbol{y}$ if $x_i < y_i$ for the first i (i goes from 1) where x_i and y_i differ. The lexicographical order between two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ of the same size is define by the lexicographical order between $\mathrm{vec}(\boldsymbol{X})$ and $\mathrm{vec}(\boldsymbol{Y})$.

Given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, the ith largest eigenvalue and the ith largest singular value of $\boldsymbol{A}$ are denoted by $\lambda_i(\boldsymbol{A})$ and $\sigma_i(\boldsymbol{A})$, respectively. The spectral radius of $\boldsymbol{A}$ is defined as $\rho(\boldsymbol{A}) = \max_i|\lambda_i(\boldsymbol{A})|$ and is less than or equal to the spectral norm, i.e., $\rho(\boldsymbol{A}) \le \|\boldsymbol{A}\|_2$. Gelfand's formula [77] states that $\rho(\boldsymbol{A}) = \lim_{k\to\infty}\|\boldsymbol{A}^k\|_2^{1/k}$. If $\boldsymbol{A}$ is square and invertible, the condition number of $\boldsymbol{A}$ is defined as $\kappa(\boldsymbol{A}) =$

$\sigma_1(\boldsymbol{A})/\sigma_n(\boldsymbol{A})$.

3.2.2 Nonlinear Orthogonal Projections

Given a non-empty set $\mathcal{C} \subset \mathbb{R}^n$, let us define the distance from a point $\boldsymbol{x} \in \mathbb{R}^n$ to $\mathcal{C}$ as

$$d(\boldsymbol{x}, \mathcal{C}) = \inf_{\boldsymbol{y} \in \mathcal{C}} \{\|\boldsymbol{y} - \boldsymbol{x}\|\}. \tag{3.2}$$

The set of all projections of $\boldsymbol{x}$ onto $\mathcal{C}$ is defined by

$$\Pi_{\mathcal{C}}(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathcal{C} \mid \|\boldsymbol{y} - \boldsymbol{x}\| = d(\boldsymbol{x}, \mathcal{C})\}. \tag{3.3}$$

It is well-known [210] that if $\mathcal{C}$ is closed, then for any $\boldsymbol{x} \in \mathbb{R}^n$, $\Pi_{\mathcal{C}}(\boldsymbol{x})$ is non-empty[3]. An orthogonal projection onto $\mathcal{C}$ is defined as $\mathcal{P}_{\mathcal{C}} : \mathbb{R}^n \to \mathcal{C}$ such that $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ is chosen as an element of $\Pi_{\mathcal{C}}(\boldsymbol{x})$ based on a prescribed scheme (e.g., based on lexicographic order). There exists a non-empty subset of $\mathbb{R}^n$ such that $\Pi_{\mathcal{C}}$ is uniquely defined, given by

$$\operatorname{singleton} \Pi_{\mathcal{C}} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \Pi_{\mathcal{C}}(\boldsymbol{x}) \text{ is singleton}\}. \tag{3.4}$$

We can now consider the differentiability of $\mathcal{P}_{\mathcal{C}}$ over $\operatorname{singleton} \Pi_{\mathcal{C}}$ as follows.

Definition 3.1 (Point-wise differentiability)**.** *The projection $\mathcal{P}_{\mathcal{C}}$ is said to be* ***dif-***

[3]In addition, if $\mathcal{C}$ is convex, then $\Pi_{\mathcal{C}}(\boldsymbol{x})$ is singleton.

ferentiable *at* $\boldsymbol{x} \in \operatorname{singleton} \Pi_{\mathcal{C}}$ *if there exists* $\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) \in \mathbb{R}^{n \times n}$ *such that*

$$\limsup_{\boldsymbol{\delta} \to \mathbf{0}} \sup_{\boldsymbol{y} \in \Pi_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})} \frac{\|\boldsymbol{y} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) - \nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})\boldsymbol{\delta}\|}{\|\boldsymbol{\delta}\|} = 0.$$

The operator $\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ *is said to be the derivative of* $\mathcal{P}_{\mathcal{C}}$ *at* $\boldsymbol{x}$.

Definition 3.2 (Point-wise Lipschitz-continuous differentiability)**.** *The projection* $\mathcal{P}_{\mathcal{C}}$ *is said to be Lipschitz-continuously differentiable at* $\boldsymbol{x} \in \operatorname{singleton} \Pi_{\mathcal{C}}$ *if* $\mathcal{P}_{\mathcal{C}}$ *is differentiable at* $\boldsymbol{x}$ *and there exist* $0 < c_1(\boldsymbol{x}) \leq \infty$ *and* $0 \leq c_2(\boldsymbol{x}) < \infty$ *such that for any* $\boldsymbol{\delta} \in \mathcal{B}(\mathbf{0}, c_1(\boldsymbol{x}))$*, we have*

$$\sup_{\boldsymbol{y} \in \Pi_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})} \|\boldsymbol{y} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) - \nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})\boldsymbol{\delta}\| \leq c_2(\boldsymbol{x})\|\boldsymbol{\delta}\|^2. \tag{3.5}$$

It is noted that the supremum in (4) implies

$$\|\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta}) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) - \nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})\boldsymbol{\delta}\| \leq c_2(\boldsymbol{x})\|\boldsymbol{\delta}\|^2$$

holds for any choice of $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})$ in $\Pi_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})$. Note that while $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ is uniquely defined for $\boldsymbol{x} \in \operatorname{singleton} \Pi_{\mathcal{C}}$, $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})$ is not since $\boldsymbol{x}+\boldsymbol{\delta}$ may not be in $\operatorname{singleton} \Pi_{\mathcal{C}}$.

Example 3.1. *Let* $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\}$ *be the unit sphere of dimension* $n-1$*. For any* $\boldsymbol{x} \neq \mathbf{0}$*, the projection onto* $\mathcal{C}$ *is uniquely given by* $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|$*. For* $\boldsymbol{x} = \mathbf{0}$*, we have* $\Pi_{\mathcal{C}}(\mathbf{0}) = \mathcal{C}$ *and* $\mathcal{P}_{\mathcal{C}}(\mathbf{0})$ *can be chosen as any point on the unit sphere. In Appendix 3.6.7, we prove that* $\mathcal{P}_{\mathcal{C}}$ *is Lipschitz-continuously differentiable at any* $\boldsymbol{x} \in \operatorname{singleton} \Pi_{\mathcal{C}} = \mathbb{R}^n \setminus \{\mathbf{0}\}$*. In particular, for any* $\boldsymbol{x} \neq \mathbf{0}$ *and* $\boldsymbol{\delta} \in \mathbb{R}^n$*, we*

have

$$\sup_{\boldsymbol{y}\in\Pi_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})}\left\|\boldsymbol{y}-\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}-\Big(\boldsymbol{I}_n-\frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|^2}\Big)\frac{\boldsymbol{\delta}}{\|\boldsymbol{x}\|}\right\|\leq\frac{2}{\|\boldsymbol{x}\|^2}\|\boldsymbol{\delta}\|^2. \tag{3.6}$$

For $\boldsymbol{\delta}\neq-\boldsymbol{x}$, $\Pi_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})=\{(\boldsymbol{x}+\boldsymbol{\delta})/\|\boldsymbol{x}+\boldsymbol{\delta}\|\}$ *is singleton and the supremum is evaluated at only one point* $\boldsymbol{y}=(\boldsymbol{x}+\boldsymbol{\delta})/\|\boldsymbol{x}+\boldsymbol{\delta}\|$. *For* $\boldsymbol{\delta}=-\boldsymbol{x}$, $\Pi_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})=\Pi_{\mathcal{C}}(\boldsymbol{0})=\mathcal{C}$ *is not singleton and the supremum is taken over the entire sphere independent of* $\boldsymbol{x}$. *In either case regardless the value of* $\boldsymbol{\delta}$, *comparing (3.6) with (3.5), we recognize the projection onto the unit sphere is Lipschitz-continuously differentiable at* $\boldsymbol{x}\in\operatorname{singleton}\Pi_{\mathcal{C}}$ *with*

$$\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})=\frac{1}{\|\boldsymbol{x}\|}\Big(\boldsymbol{I}_n-\frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|^2}\Big),$$
$$c_1(\boldsymbol{x})=\infty,\quad c_2(\boldsymbol{x})=\frac{2}{\|\boldsymbol{x}\|^2}.$$

In 1984, Foote [67] showed that if $\mathcal{C}$ is a C^k $(k\geq 2)$ submanifold of $\mathbb{R}^n$, then $\mathcal{C}$ has a neighborhood $\mathcal{E}$ such that $\mathcal{E}\subseteq\operatorname{singleton}\Pi_{\mathcal{C}}$ and the projection $\mathcal{P}_{\mathcal{C}}$ restricted to $\mathcal{E}$ is a C^{k-1} mapping. Later on, Dudek and Holly [59] proved the derivative $\nabla\mathcal{P}_{\mathcal{C}}$ is a linear map to the tangent bundle of $\mathcal{C}$ and more importantly, for any $\boldsymbol{x}^*\in\mathcal{C}$, $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)$ is the (linear) orthogonal projection onto the tangent space to $\mathcal{C}$ at $\boldsymbol{x}^*$. Recently, a local version of this result has been proposed by Lewis and Malick [129]:

Proposition 3.1. *(Rephrased from Lemma 4 in [129]) Assume* $\mathcal{C}$ *is a* C^k *(*$k\geq 2$*) manifold around a point* $\boldsymbol{x}^*\in\mathcal{C}$. *Denote the tangent space to* $\mathcal{C}$ *at* $\boldsymbol{x}^*$ *by* $T_{\boldsymbol{x}^*}(\mathcal{C})$.

Algorithm 3.1 Projected Gradient Descent (PGD)

Input: f, $\mathcal{C}$, η, $\boldsymbol{x}^{(0)}$
Output: $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$

1: **for** $k = 0, 1, \ldots$ **do**
2: $\quad \boldsymbol{z}_{\eta}^{(k)} = \boldsymbol{x}^{(k)} - \eta \boldsymbol{A}^{\top}\big(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b}\big)$
3: $\quad \boldsymbol{x}^{(k+1)} = \mathcal{P}_{\mathcal{C}}\big(\boldsymbol{z}_{\eta}^{(k)}\big)$

Then, the set of projections $\Pi_{\mathcal{C}}$ is (locally) singleton around $\boldsymbol{x}^*$. Moreover, $\mathcal{P}_{\mathcal{C}}$ is a C^{k-1} mapping around $\boldsymbol{x}^*$ and

$$\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \mathcal{P}_{T_{\boldsymbol{x}^*}(\mathcal{C})}, \tag{3.7}$$

where $\mathcal{P}_{T_{\boldsymbol{x}^*}(\mathcal{C})}$ is the orthogonal projection onto $T_{\boldsymbol{x}^*}(\mathcal{C})$.

Further works on the uniqueness and regularity of $\mathcal{P}_{\mathcal{C}}$ can also be found in [4, 6, 127, 173]. We note that the assumption $\mathcal{C}$ is a C^2 manifold around $\boldsymbol{x}^*$ requires the existence of a neighborhood of $\boldsymbol{x}^*$ in which $\mathcal{P}_{\mathcal{C}}$ is uniformly differentiable. Our result in this chapter, while strongly motivated by the aforementioned results, only requires $\mathcal{C}$ to be differentiable at two points (see Theorem 3.1).

3.2.3 Stationary Points of (3.1)

We defined the (Lipschitz-continuous) differentiability of the projection $\mathcal{P}_{\mathcal{C}}$ at a point in $\mathcal{C}$. We are now in position to define stationary points of (3.1) as those where the gradient of the objective function on the constraint set vanishes [3]:

Definition 3.3. *$\boldsymbol{x}^* \in \mathcal{C}$ is a **stationary point** of (3.1) if $\mathcal{P}_{\mathcal{C}}$ is differentiable at*

$\boldsymbol{x}^*$ *and*

$$\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)\boldsymbol{A}^\top\big(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\big) = \boldsymbol{0}. \tag{3.8}$$

Assume in addition that $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at $\boldsymbol{x}^*$ with constants $c_1(\boldsymbol{x}^*)$ and $c_2(\boldsymbol{x}^*)$. Then $\boldsymbol{x}^*$ is called a ***Lipschitz stationary point*** *of (3.1) with constants $c_1(\boldsymbol{x}^*)$ and $c_2(\boldsymbol{x}^*)$.*

Similar to unconstrained optimization, stationary points in Definition 3.3 can be local minimizers, local maximizers, or saddle points of the constrained problem (3.1).

3.2.4 Projected Gradient Descent

Algorithm 3.1 describes the projected gradient descent algorithm for solving (3.1). Starting at some $\boldsymbol{x}^{(0)} \in \mathcal{C}$, the algorithm iteratively updates the current value by *(i)* taking a step in the opposite direction of the gradient and *(ii)* projecting the result back onto $\mathcal{C}$, i.e.,

$$\boldsymbol{x}^{(k+1)} = \mathcal{P}_{\mathcal{C}}\Big(\boldsymbol{x}^{(k)} - \eta\boldsymbol{A}^\top\big(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b}\big)\Big), \tag{3.9}$$

where $\eta > 0$ is a fixed step size.

Definition 3.4. *$\boldsymbol{x}^*$ is a **fixed point** of Algorithm 3.1 with step size $\eta > 0$ if*

$$\boldsymbol{x}^* = \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* - \eta \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})). \tag{3.10}$$

Lemma 3.1. *If $\boldsymbol{x}^*$ is a fixed point of Algorithm 3.1 with some step size $\eta > 0$ and $\mathcal{P}_{\mathcal{C}}$ is differentiable at $\boldsymbol{x}^*$, then $\boldsymbol{x}^*$ is a stationary point of (3.1).*

The proof of Lemma 3.1 is given in Appendix 3.6.1.

3.3 Local Convergence Analysis

In this section, we present the key contribution of this work, namely, a local convergence analysis of projected gradient descent for constrained least squares. Specifically, **our goal** is to establish the following results: *(i)* a closed-form expression of the exact asymptotic rate of convergence, *(ii)* a bound on the number of iterations needed to reach a certain level of accuracy, and *(iii)* a region of convergence. Figure 3.1 illustrates the key idea in our analysis. In order to establish the local linear convergence of Algorithm 3.1 to its fixed point $\boldsymbol{x}^*$, we require the Lipschitz-continuous differentiability of $\mathcal{P}_{\mathcal{C}}$ at $\boldsymbol{x}^*$ and at $\boldsymbol{z}_\eta^* = \boldsymbol{x}^* - \eta \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$. These properties enables us to approximate each projected gradient descent update by a linear operator on the error vector (i.e., the difference between $\boldsymbol{x}^{(k)}$ and $\boldsymbol{x}^*$). Then, under the additional assumption that this linear operator is a contraction mapping and the initialization $\boldsymbol{x}^{(0)}$ is sufficiently close to $\boldsymbol{x}^*$, we show that the gradient step and the projection step remain inside the Lipschitz-continuous differentiability

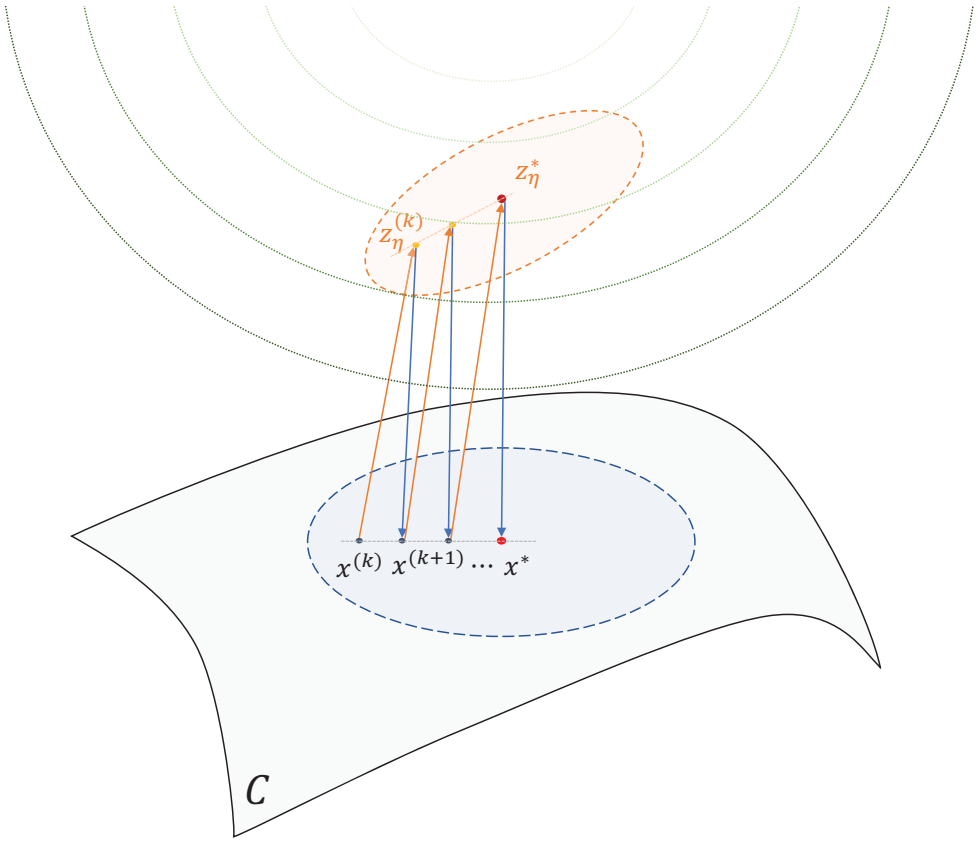


Figure 3.1: Illustration of convergence of projected gradient descent to a fixed point $\boldsymbol{x}^*$. In order to guarantee linear convergence, Theorem 3.1 requires $\mathcal{P}_\mathcal{C}$ to be Lipschitz-continuously differentiable at both $\boldsymbol{x}^{(k)}$ and $\boldsymbol{z}_\eta^{(k)} = \boldsymbol{x}^{(k)} - \eta \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b})$. Moreover, the condition $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\| < \min\{c_1(\boldsymbol{x}^*)/\kappa(\boldsymbol{Q}), c_1(\boldsymbol{z}_\eta^*)/(\kappa(\boldsymbol{Q})u_\eta)\}$ from (3.13) ensures that $\boldsymbol{x}^{(k)}$ remains inside $\mathcal{B}(\boldsymbol{x}^*, c_1(\boldsymbol{x}^*))$ (blue dashed ellipse) and $\boldsymbol{z}_\eta^{(k)}$ = remains inside $\mathcal{B}(\boldsymbol{z}_\eta^*, c_1(\boldsymbol{z}_\eta^*))$ (orange dashed ellipse).

regions of $\boldsymbol{x}^*$ (i.e., $\mathcal{B}(\boldsymbol{x}^*, c_1(\boldsymbol{x}^*))$) and $\boldsymbol{z}_\eta^*$ (i.e., $\mathcal{B}(\boldsymbol{z}_\eta^*, c_1(\boldsymbol{z}_\eta^*))$, respectively).

3.3.1 Main Results

In this following, we state our main result in Theorem 3.1, followed by further insights into the convergence results.

Theorem 3.1. *Suppose $\boldsymbol{x}^*$ is a fixed point of Algorithm 3.1 with step size $\eta > 0$ such that the following conditions hold:*

1. *$\mathcal{P}_\mathcal{C}$ is Lipschitz-continuously differentiable at both the fixed point $\boldsymbol{x}^*$ and at the gradient step taken from the fixed point*

$$\boldsymbol{z}_\eta^* = \boldsymbol{x}^* - \eta \boldsymbol{A}^\top \big(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b} \big), \tag{3.11}$$

   with the corresponding matrices $\nabla \mathcal{P}_\mathcal{C}(\boldsymbol{x}^*)$, $\nabla \mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*)$, and constants $c_1(\boldsymbol{x}^*)$, $\boldsymbol{c}_2(\boldsymbol{x}^*)$, $c_1(\boldsymbol{z}_\eta^*)$, and $c_2(\boldsymbol{z}_\eta^*)$.

2. *The matrix*

$$\boldsymbol{H} = \nabla \mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*)(\boldsymbol{I}_n - \eta \boldsymbol{A}^\top \boldsymbol{A}) \nabla \mathcal{P}_\mathcal{C}(\boldsymbol{x}^*) \tag{3.12}$$

   admits an eigendecomposition $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$, where $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is an invertible matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal entries are strictly less than 1 in magnitude, i.e., $\rho(\boldsymbol{H}) = \|\boldsymbol{\Lambda}\|_2 < 1$.

3. The initial iterate $\boldsymbol{x}^{(0)}$ *satisfies*

$$\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\| < \min\Big\{\frac{c_1(\boldsymbol{x}^*)}{\kappa(\boldsymbol{Q})}, \frac{c_1(\boldsymbol{z}_\eta^*)}{\kappa(\boldsymbol{Q})u_\eta}, \frac{1-\rho(\boldsymbol{H})}{q}\Big\}, \tag{3.13}$$

where

$$u_\eta = \|\boldsymbol{I}_n - \eta \boldsymbol{A}^\top \boldsymbol{A}\|_2 \tag{3.14}$$

and

$$q = \kappa^2(\boldsymbol{Q})u_\eta\big(c_2(\boldsymbol{z}_\eta^*)u_\eta + \|\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)\|_2 c_2(\boldsymbol{x}^*)\big). \tag{3.15}$$

Let $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ *be the vector sequence generated by the PGD update in (3.9). Then, for any* $0 < \epsilon < 1$*, we have* $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| \leq \epsilon \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|$ *for all*

$$k \geq \frac{\log(1/\epsilon) + \log(\kappa(\boldsymbol{Q}))}{\log(1/\rho(\boldsymbol{H}))} + c_3, \tag{3.16}$$

where $c_3 > 0$*, given explicitly in Lemma 3.4, is independent of* ϵ*. Algorithm 3.1 is said to converge locally to* $\boldsymbol{x}^*$ *at an* ***asymptotic linear rate*** $\rho(\boldsymbol{H})$ *with the* ***region of linear convergence*** *given by (3.13).*

Theorem 3.1 states the sufficient conditions for asymptotic linear convergence of Algorithm 3.1. In addition, the theorem establishes the asymptotic rate as the spectral radius of the matrix $\boldsymbol{H}$ and bounds the number of iterations needed to reach ϵ-accuracy. The proof of Theorem 3.1 is given in Subsection 3.3.2. It is note-

worthy that in the RHS of (3.16), the first term corresponds to linear convergence in the asymptotic regime and the second term corresponds to nonlinear convergence behavior at the early stage. We will revisit this point when we introduce Lemma 3.4.

Remark 3.1. *When* $\boldsymbol{H}$ *is symmetric, its eigendecomposition exists and can be represented as*

$$\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top,$$

where $\boldsymbol{Q}$ *is an orthogonal matrix with* $\kappa(\boldsymbol{Q}) = 1$.

Next, we study a special case of Theorem 3.1 in which

$$\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*) = \nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \mathcal{P}_{T_{\boldsymbol{x}^*}(\mathcal{C})} = \boldsymbol{U}_{\boldsymbol{x}^*}\boldsymbol{U}_{\boldsymbol{x}^*}^\top, \tag{3.17}$$

where $\boldsymbol{U}_{\boldsymbol{x}^*} \in \mathbb{R}^{n\times d}$ $(d \leq n)$ is the matrix whose columns provide an orthonormal basis for the tangent space to $\mathcal{C}$ at $\boldsymbol{x}^*$. A typical example in which (3.17) holds is when *(i)* $\mathcal{C}$ is a C^2 d-dimensional submanifold around $\boldsymbol{x}^*$; and *(ii)* $\boldsymbol{z}_\eta^* = \boldsymbol{x}^*$. The first condition *(i)* stems from Proposition 3.1 in order to guarantee $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \mathcal{P}_{T_{\boldsymbol{x}^*}(\mathcal{C})}$. The second condition *(ii)* is equivalent to $\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{0}$, which means $\boldsymbol{x}^*$ is also a stationary point of the unconstrained problem. Conveniently, this coincidence eliminates the task of characterizing the projection $\mathcal{P}_{\mathcal{C}}$ and its derivative $\nabla\mathcal{P}_{\mathcal{C}}$ at a point outside $\mathcal{C}$, which can be a challenging task in many problems.

Corollary 3.1. *Consider the same setting as in Theorem 3.1 with the additional assumption that (3.17) holds. If* $(\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}$ *has full rank and*

$$0 < \eta < \frac{2}{\|\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}\|_2^2}, \tag{3.18}$$

then Algorithm 3.1 with fixed step size η *converges locally to* $\boldsymbol{x}^*$ *at an asymptotic linear rate*

$$\rho(\boldsymbol{H}) = \max\{|1 - \eta\lambda_1|, |1 - \eta\lambda_d|\}, \tag{3.19}$$

where λ_1 *and* λ_d *are the largest and smallest eigenvalues of* $(\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}$*, respectively. The region of linear convergence is given by*

$$\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\| < \min\Big\{c_1(\boldsymbol{x}^*), \frac{c_1(\boldsymbol{z}_\eta^*)}{u_\eta}, \frac{1 - \rho(\boldsymbol{H})}{u_\eta c_2(\boldsymbol{x}^*) + u_\eta^2 c_2(\boldsymbol{z}_\eta^*)}\Big\}, \tag{3.20}$$

where u_η *is given by (3.14).*

The proof of Corollary 3.1 is given in Appendix 3.6.2.

Remark 3.2. *Recall that* u_η *defined in (3.14) is also the asymptotic linear rate of gradient descent for the unconstrained least squares [167], i.e.,*

$$u_\eta = \max\{|1 - \eta\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})|, |1 - \eta\lambda_n(\boldsymbol{A}^\top\boldsymbol{A})|\}.$$

Since $\boldsymbol{U}_{\boldsymbol{x}^*}$ *is a semi-orthogonal matrix, the eigenvalues of* $\boldsymbol{U}_{\boldsymbol{x}^*}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}$ *interlace with those of* $\boldsymbol{A}^\top\boldsymbol{A}$ *[101], which in turns implies* $\lambda_n(\boldsymbol{A}^\top\boldsymbol{A}) \le \lambda_d \le \lambda_1 \le \lambda_1(\boldsymbol{A}^\top\boldsymbol{A})$.

Thus, one can show that for $\eta < 2/\|\boldsymbol{A}\|_2^2$,

$$\rho(\boldsymbol{H}) \leq u_\eta \leq 1, \tag{3.21}$$

with the equality $u_\eta = 1$ *holding if and only if* $\boldsymbol{A}^\top \boldsymbol{A}$ *is singular. Interestingly, (3.21) implies the presence of the constraint in this case helps accelerate the convergence of gradient descent to* $\boldsymbol{x}^*$.

3.3.2 Proof of Theorem 3.1

This section presents the proof of Theorem 3.1. Our key ideas are: *(1)* using the Lipschitz-continuous differentiability of $\mathcal{P}_\mathcal{C}$ at $\boldsymbol{x}^*$ and at $\boldsymbol{z}_\eta^*$ to establish a recursive relation on the error vector $\boldsymbol{\delta}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^*$, *(2)* performing a change of basis $\tilde{\boldsymbol{\delta}}^{(k)} = \boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(k)}$ to establish an asymptotically-linear quadratic system dynamic that upper-bounds the norm of the transformed error vector, *(3)* applying the result on the convergence of an asymptotically-linear quadratic difference equation in [216] to obtain the number of iterations required for $\|\tilde{\boldsymbol{\delta}}^{(k)}\| \leq \tilde{\epsilon}\|\tilde{\boldsymbol{\delta}}^{(0)}\|$, and *(4)* converting the convergence result on the transformed error $\|\tilde{\boldsymbol{\delta}}^{(k)}\|$ to the convergence result on the original error $\|\boldsymbol{\delta}^{(k)}\|$. In the following, we provide the complete proof, with some details deferred to the appendix.

Step 1: Let us define the error vector of Algorithm 3.1 as $\boldsymbol{\delta}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^*$, for $k \in \mathbb{N}$. Using this definition of the error vector, we can replace $\boldsymbol{x}^{(k)} = \boldsymbol{x}^* + \boldsymbol{\delta}^{(k)}$

and $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^* + \boldsymbol{\delta}^{(k+1)}$ into (3.9) to obtain an equivalent update on the error vector

$$\boldsymbol{\delta}^{(k+1)} = \mathcal{P}_{\mathcal{C}}\Big(\boldsymbol{x}^* + \boldsymbol{\delta}^{(k)} - \eta \boldsymbol{A}^\top\big(\boldsymbol{A}(\boldsymbol{x}^* + \boldsymbol{\delta}^{(k)}) - \boldsymbol{b}\big)\Big) - \boldsymbol{x}^*. \tag{3.22}$$

Based on the definition of $\boldsymbol{z}_\eta^*$ in (3.11) and the fact that $\boldsymbol{x}^*$ is a fixed point of the algorithm (see (3.10)), i.e., $\boldsymbol{x}^* = \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)$, we can rewrite (3.22) as

$$\boldsymbol{\delta}^{(k+1)} = \mathcal{P}_{\mathcal{C}}\Big(\boldsymbol{z}_\eta^* + (\boldsymbol{I} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}\Big) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*). \tag{3.23}$$

We are now in position to analyze the error update as a fixed-point iteration: $\boldsymbol{\delta}^{(k+1)} = \boldsymbol{f}(\boldsymbol{\delta}^{(k)})$, where $\boldsymbol{f}(\boldsymbol{\delta}) = \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^* + (\boldsymbol{I} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)$. The following lemma provides a recursive equation on the error vector that is in the form of an asymptotically-linear quadratic system dynamic:

Lemma 3.2. *Recall* $\boldsymbol{H} = \nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)(\boldsymbol{I}_n - \eta \boldsymbol{A}^\top \boldsymbol{A}) \nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)$. *If the error vector at the* k*-th iteration satisfies*

$$\|\boldsymbol{\delta}^{(k)}\| < \min\Big\{c_1(\boldsymbol{x}^*), \frac{c_1(\boldsymbol{z}_\eta^*)}{u_\eta}\Big\}, \tag{3.24}$$

then the error vector at the $k+1$*-th iteration satisfies*

$$\boldsymbol{\delta}^{(k+1)} = \boldsymbol{H}\boldsymbol{\delta}^{(k)} + \boldsymbol{q}_2(\boldsymbol{\delta}^{(k)}), \tag{3.25}$$

where $\boldsymbol{q}_2 : \mathbb{R}^n \to \mathbb{R}^n$ *is the residual such that*

$$\|\boldsymbol{q}_2(\boldsymbol{\delta}^{(k)})\| \leq \big(\|\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)\|_2 u_\eta c_2(\boldsymbol{x}^*) + c_2(\boldsymbol{z}_\eta^*) u_\eta^2\big)\|\boldsymbol{\delta}^{(k)}\|^2. \tag{3.26}$$

The proof of Lemma 2 is given in Appendix 3.6.3. Given the nonlinear difference equation of form (3.25), we proceed with characterizing the convergence of the error sequence $\{\boldsymbol{\delta}^{(k)}\}_{k=0}^{\infty}$.

Remark 3.3. *Dating back to 1964, Polyak [166] studied the convergence of nonlinear difference equations of form*

$$\boldsymbol{a}^{(k+1)} = \boldsymbol{T}(\boldsymbol{a}^{(k)}) + \boldsymbol{q}(\boldsymbol{a}^{(k)}), \quad \text{for } k \in \mathbb{N}, \tag{3.27}$$

where $\boldsymbol{a}^{(0)} \in \mathbb{R}^n$, $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^n$ *is a linear operator, and* $\boldsymbol{q} : \mathbb{R}^n \to \mathbb{R}^n$ *satisfies* $\lim_{t\to 0} \sup_{\|\boldsymbol{a}\|\leq t} \|\boldsymbol{q}(\boldsymbol{a})\| / \|\boldsymbol{a}\| = 0$. *The author showed that if the operator* $\boldsymbol{T}$ *satisfies* $\|\boldsymbol{T}^k\|_2 \leq c(\zeta)(\rho+\zeta)^k$, *for some* $\rho < 1$ *and arbitrarily small* $\zeta > 0$, *then* $\{\boldsymbol{a}^{(k)}\}_{k=0}^{\infty}$ *approaches zero with sufficiently small* $\|\boldsymbol{a}^{(0)}\|$*:*

$$\|\boldsymbol{a}^{(k)}\| \leq C(\zeta)\|\boldsymbol{a}^{(0)}\|(\rho+\zeta)^k. \tag{3.28}$$

Here $c(\zeta)$ *and* $C(\zeta)$ *are unknown constants that could grow to infinity as* $\zeta \to 0$. *Applying this result to (3.25) with* $\boldsymbol{a}^{(k)} = \boldsymbol{\delta}^{(k)}$ *and* $\boldsymbol{T} = \boldsymbol{H}$, *one can show that the error vector of Algorithm 3.1 converges to* $\mathbf{0}$ *with the asymptotic linear rate* $\rho(\boldsymbol{H})$, *provided that* $\rho(\boldsymbol{H}) < 1$ *and* $\|\boldsymbol{\delta}^{(0)}\|$ *is sufficiently small. However, we note that the proof of (3.28) in [166] is adapted from a more general result on the*

stability of differential equations in [14]. This technique can not provide the precise control of the ROC and the number of iterations required to reach a certain accuracy (i.e., how small $\|\boldsymbol{a}^{(0)}\|$ is as well as how large the factor $C(\zeta)$ is) needed for our convergence analysis of PGD. Alternatively, we utilize our previous result in [216] that eliminates the dependence on ζ in the expression of the linear rate, at the cost of an additional assumption on the diagonalizability of $\boldsymbol{H}$.[4] *Additionally, our approach offers explicit expressions of the ROC and the number of required iterations (as in (3.13) and (3.16), respectively).*

Step 2: Our approach for analyzing the convergence of the nonlinear difference equation (24) is to leverage the eigendecomposition $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$ and consider the transformed error vector as follows.

Lemma 3.3. *Let $\tilde{\boldsymbol{\delta}}^{(k)} = \boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(k)}$ be the transformed error vector. If (3.13) holds and the spectral radius of $\boldsymbol{H}$ is strictly less than 1, i.e., $\rho(\boldsymbol{H}) < 1$, then, for all $k \in \mathbb{N}$, we have*

$$\tilde{\boldsymbol{\delta}}^{(k+1)} = \boldsymbol{\Lambda}\tilde{\boldsymbol{\delta}}^{(k)} + \boldsymbol{q}_3(\tilde{\boldsymbol{\delta}}^{(k)}), \tag{3.29}$$

where the residual $\boldsymbol{q}_3(\tilde{\boldsymbol{\delta}}^{(k)}) = \boldsymbol{Q}^{-1}\boldsymbol{q}_2(\boldsymbol{Q}\tilde{\boldsymbol{\delta}}^{(k)})$ satisfies $\|\boldsymbol{q}_3(\tilde{\boldsymbol{\delta}}^{(k)})\| \leq (q/\|\boldsymbol{Q}^{-1}\|_2)\|\tilde{\boldsymbol{\delta}}^{(k)}\|^2$ for q given in (3.15).

The proof of Lemma 3.3 is given in Appendix 3.6.4. Taking the norms of both

[4] In particular, the bound in (3.16) suggests $\|\boldsymbol{a}^{(k)}\| \leq C\|\boldsymbol{a}^{(0)}\|\rho^k$, for constant $C = \rho\kappa(\boldsymbol{Q})e^{c_3}$, which is tighter than (3.28).

sides of (3.29) and applying the triangle inequality, we obtain

$$\|\tilde{\boldsymbol{\delta}}^{(k+1)}\| \leq \rho(\boldsymbol{H})\|\tilde{\boldsymbol{\delta}}^{(k)}\| + \frac{q}{\|\boldsymbol{Q}^{-1}\|_2}\|\tilde{\boldsymbol{\delta}}^{(k)}\|^2. \tag{3.30}$$

This inequality, holding for all $k \in \mathbb{N}$, is the key to the convergence of the transformed error sequence in the next step.

Step 3: If we replace the inequality symbol in (3.30) by the equality symbol, then we obtain an asymptotically-linear quadratic difference equation whose convergence is studied in [216]. Indeed, the following lemma states that the norm of the transformed error vector is governed by this asymptotically-linear quadratic system dynamic:

Lemma 3.4. *Assume the same setting as Lemma 3.3. Then, for any desired accuracy* $0 < \tilde{\epsilon} < 1$*, we have* $\|\tilde{\boldsymbol{\delta}}^{(k)}\| \leq \tilde{\epsilon}\|\tilde{\boldsymbol{\delta}}^{(0)}\|$ *for all*

$$k \geq \frac{\log(1/\tilde{\epsilon})}{\log(1/\rho(\boldsymbol{H}))} + c_3(\rho(\boldsymbol{H}), \tau), \tag{3.31}$$

where $\tau = q\|\tilde{\boldsymbol{\delta}}^{(0)}\|/\|\boldsymbol{Q}^{-1}\|_2/(1-\rho(\boldsymbol{H})) \in (0,1)$ *and*

$$\begin{aligned} c_3(\rho,\tau) = \; & \frac{E_1\Big(\log \frac{1}{\rho+\tau(1-\rho)}\Big) - E_1\Big(\log \frac{1}{\rho}\Big)}{\rho \log(1/\rho)} \\ & + \frac{1}{2\rho}\log\left(\frac{\log(1/\rho)}{\log\big(1/(\rho+\tau(1-\rho))\big)}\right) + 1, \end{aligned} \tag{3.32}$$

for $E_1(t) = \int_t^\infty \frac{e^{-z}}{z} dz$ *being the exponential integral [2].*

The proof of Lemma 3.4 is given in Appendix 3.6.5.

Step 4: Finally, we show the convergence of $\|\boldsymbol{\delta}^{(k)}\|$ based on the convergence of $\|\tilde{\boldsymbol{\delta}}^{(k)}\|$. From (3.31), substituting $\tilde{\epsilon} = \epsilon/\kappa(\boldsymbol{Q})$ and identifying c_3 as $c_3(\rho(\boldsymbol{H}), \tau)$, we obtain (3.16). Thus, it remains to prove that the accuracy on the transformed error vector $\|\tilde{\boldsymbol{\delta}}^{(k)}\| \leq \tilde{\epsilon}\|\tilde{\boldsymbol{\delta}}^{(0)}\|$ is sufficient for the accuracy on the original error vector $\|\boldsymbol{\delta}^{(k)}\| \leq \epsilon\|\boldsymbol{\delta}^{(0)}\|$. Indeed, given

$$\|\tilde{\boldsymbol{\delta}}^{(k)}\| \leq \tilde{\epsilon}\|\tilde{\boldsymbol{\delta}}^{(0)}\| = \frac{\epsilon}{\|\boldsymbol{Q}\|_2\|\boldsymbol{Q}^{-1}\|_2}\|\tilde{\boldsymbol{\delta}}^{(0)}\|,$$

we have

$$\begin{aligned}\|\boldsymbol{\delta}^{(k)}\| = \|\boldsymbol{Q}\tilde{\boldsymbol{\delta}}^{(k)}\| &\leq \|\boldsymbol{Q}\|_2\|\tilde{\boldsymbol{\delta}}^{(k)}\| \\ &\leq \|\boldsymbol{Q}\|_2\frac{\epsilon}{\|\boldsymbol{Q}\|_2\|\boldsymbol{Q}^{-1}\|_2}\|\tilde{\boldsymbol{\delta}}^{(0)}\| \\ &= \frac{\epsilon}{\|\boldsymbol{Q}^{-1}\|_2}\|\tilde{\boldsymbol{\delta}}^{(0)}\| \leq \epsilon\|\boldsymbol{\delta}^{(0)}\|,\end{aligned}$$

where the last inequality stems from $\|\tilde{\boldsymbol{\delta}}^{(0)}\| = \|\boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(0)}\| \leq \|\boldsymbol{Q}^{-1}\|_2\|\boldsymbol{\delta}^{(0)}\|$. This completes our proof of Theorem 3.1.

3.4 Applications

In this section, we demonstrate the application of our proposed framework to a collection of well-known problems in machine learning and signal processing. The constraint sets in these problems vary from as simple as an affine subspace (A) and a sphere (C) to more complex algebraic varieties such as the s-sparse vector set (B)

and the low-rank matrix set (D). We consider both problems with known convergence rate results and problems for which the rate is unavailable. The former allows us to verify the correctness of our analysis against the known rate results, while for the latter numerical experiments are used to verify the rate. Additionally, we illustrate how ROC can be obtained for each problem. Due to space limitation, we restrict the illustration of our framework to the four aforementioned applications. While we believe that additional applications can be considered (see the potential applications of our framework in Section V), such applications may require a more elaborate development. Our goal in this section is to offer a recipe for analyzing the convergence of PGD for different applications using the proposed framework. Table 3.1 describes the steps we follow to obtain the asymptotic linear rate and the region of linear convergence in each application. Table 3.2 summarizes our local convergence results on the four problems presented in this section. The detailed analysis is given below.

3.4.1 Linear Equality-Constrained Least Squares

As a sanity check, we start with a simple example of the so-called linear equality-constrained least squares (LECLS)

$$\boxed{\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2 \quad \text{s.t. } \boldsymbol{C}\boldsymbol{x}=\boldsymbol{d},} \tag{3.33}$$

Table 3.1: General recipe for local convergence analysis.

Step 1:	Identify $\boldsymbol{A}$, $\boldsymbol{b}$, $\mathcal{C}$, and $\mathcal{P}_{\mathcal{C}}$.
Step 2:	Establish the conditions for $\boldsymbol{x}^* \in \mathcal{C}$ to be a **Lipschitz stationary point** of (3.1). In particular, *(i)* $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at every $\boldsymbol{x}^*$ with $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)$, $c_1(\boldsymbol{x}^*)$, and $c_2(\boldsymbol{x}^*)$; and *(ii)* the stationarity equation (3.8) holds.
Step 3:	Establish the conditions for $\eta > 0$ such that *(i)* $\boldsymbol{x}^*$ is a **fixed point** of Algorithm 3.1 with step size η, i.e., $\boldsymbol{x}^* = \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)$, for $\boldsymbol{z}_\eta^* = \boldsymbol{x}^* - \eta\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$; and *(ii)* $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at $\boldsymbol{z}_\eta^*$ with $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)$, $c_1(\boldsymbol{z}_\eta^*)$, and $c_2(\boldsymbol{z}_\eta^*)$.
Step 4:	Determine the **asymptotic linear rate** ρ as the spectral radius of $\boldsymbol{H}$ given by (3.12). (If $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*) = \nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)$, (3.19) can be used instead.)
Step 5:	Establish the conditions for $\rho < 1$, which guarantees local linear convergence. Thereby, combine these conditions with the previous conditions obtained from Steps 2 and 3.
Step 6:	If $\boldsymbol{H}$ is diagonalizable, determine the **region of linear convergence** given by (3.13). (If $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*) = \nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)$, (3.20) can be used instead.)

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{C} \in \mathbb{R}^{p\times n}$, and $\boldsymbol{d} \in \mathbb{R}^p$. In addition, we assume that $p < n$ and $\boldsymbol{C}$ has linearly independent rows. The LECLS problem finds application in a wide range of areas such as linear-phase system identification [96], antenna array processing [53], and adaptive array processing [69]. While this problem can be solved efficiently using the method of Lagrange multipliers [177] or the method of weighting [209], we limit our interest to using PGD to solve (3.33) to demonstrate the applicability of our analysis. In the literature, this algorithm is referred to as the projected Landweber iteration [17, 60, 110, 143]. While these works provide bounds on the linear convergence of PGD for different variants of linear equality-constrained problems, we have not found any closed-form expression of the asymptotic rate of linear convergence.

Table 3.2: Summary of local convergence analysis for four problems: linear equality-constrained least squares (Sec. 3.4.1), sparse recovery (Sec. 3.4.2), least squares with a unit norm constraint (Sec. 3.4.3), and matrix completion (Sec. 3.4.4). In the second row, $\boldsymbol{v}^* = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$. In the third row, $\boldsymbol{K} = (\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}$. We refer the reader to each of the corresponding sections for further details.

Problem formulation	**Condition(s) for linear convergence**	**Asymptotic rate of convergence ρ**	**Region of convergence**
$\min \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$ s.t. $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}$	$\begin{cases} \boldsymbol{K} = (\boldsymbol{A}\boldsymbol{V}_C^\perp)^\top \boldsymbol{A}\boldsymbol{V}_C^\perp \text{ has full rank} \\ 0 < \eta < 2/\|\boldsymbol{A}\boldsymbol{V}_C^\perp\|_2^2 \end{cases}$	$\max\{\lvert 1 - \eta\lambda_1(\boldsymbol{K})\rvert, \lvert 1 - \eta\lambda_{n-p}(\boldsymbol{K})\rvert\}$	$\|\boldsymbol{x} - \boldsymbol{x}^*\| < \infty$
$\min \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$ s.t. $\|\boldsymbol{x}\|_0 \le s$	$\begin{cases} \boldsymbol{K} = (\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*} \text{ has full rank} \\ 0 < \eta < \min\{\frac{2}{\|\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}\|_2^2}, \frac{\lvert x^*_{[s]}\rvert}{\|\boldsymbol{v}^*\|_\infty}\} \end{cases}$	$\max\{\lvert 1 - \eta\lambda_1(\boldsymbol{K})\rvert, \lvert 1 - \eta\lambda_s(\boldsymbol{K})\rvert\}$	$\|\boldsymbol{x} - \boldsymbol{x}^*\| < \min\{\frac{\lvert x^*_{[s]}\rvert}{\sqrt{2}}, \frac{\lvert x^*_{[s]}\rvert - \eta\|\boldsymbol{v}^*\|_\infty}{\sqrt{2}\|\boldsymbol{I}_n - \eta\boldsymbol{A}^\top\boldsymbol{A}\|_2}\}$
$\min \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$ s.t. $\|\boldsymbol{x}\| = 1$	$\begin{cases} 0 < \eta < \infty \quad \text{if } \gamma \le -\lambda_1(\boldsymbol{K}) \\ 0 < \eta < \frac{2}{\gamma + \lambda_1(\boldsymbol{K})} \quad \text{o.t.w.} \end{cases}$	$\frac{1}{1-\eta\gamma}\max\{\lvert 1 - \eta\lambda_1(\boldsymbol{K})\rvert, \lvert 1 - \eta\lambda_{n-1}(\boldsymbol{K})\rvert\}$	$\|\boldsymbol{x} - \boldsymbol{x}^*\| \le \frac{1-\rho}{2(t^2+t)},\ t = \frac{\|\boldsymbol{I}_n - \eta\boldsymbol{A}^\top\boldsymbol{A}\|_2}{1-\eta\gamma}$
$\min \frac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{X} - \boldsymbol{M})\|_F^2$ s.t. $\operatorname{rank}(\boldsymbol{X}) \le r$	$\begin{cases} \boldsymbol{K} = \boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp \text{ has full rank} \\ 0 < \eta < \frac{2}{\|\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp\|_2} \end{cases}$	$\max\{\lvert 1 - \eta\lambda_1(\boldsymbol{K})\rvert, \lvert 1 - \eta\lambda_{r(m+n-r)}(\boldsymbol{K})\rvert\}$	$\|\boldsymbol{X} - \boldsymbol{X}^*\|_F \le \frac{(1-\rho)\sigma_r(\boldsymbol{X}^*)}{8(1+\sqrt{2})}$

Step 1: In this example, $\boldsymbol{A}$ and $\boldsymbol{b}$ are given explicitly in (3.33). The constraint set $\mathcal{C}$ is the closed convex affine subspace

$$\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}\}.$$

The orthogonal projection onto this subspace is given in a closed-form expression as $\mathcal{P}_\mathcal{C}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\boldsymbol{x} - \boldsymbol{d})$, for all $\boldsymbol{x} \in \mathbb{R}^n$ [151]. Since $\boldsymbol{C}$ has full row rank, it admits a compact singular value decomposition (SVD) $\boldsymbol{C} = \boldsymbol{U}_C \boldsymbol{\Sigma}_C \boldsymbol{V}_C^\top$, where $\boldsymbol{\Sigma}_C \in \mathbb{R}^{p\times p}$ is a diagonal matrix with positive diagonal entries, $\boldsymbol{U}_C \in \mathbb{R}^{p\times p}$ and $\boldsymbol{V}_C \in \mathbb{R}^{n\times p}$ satisfy $\boldsymbol{U}_C^\top \boldsymbol{U}_C = \boldsymbol{V}_C^\top \boldsymbol{V}_C = \boldsymbol{I}_p$. Denote $\boldsymbol{V}_C^\perp \in \mathbb{R}^{n\times(n-p)}$ the orthogonal complement of $\boldsymbol{V}_C$, i.e., $\boldsymbol{V}_C^\perp (\boldsymbol{V}_C^\perp)^\top = \boldsymbol{I}_n - \boldsymbol{V}_C \boldsymbol{V}_C^\top$ and $(\boldsymbol{V}_C^\perp)^\top \boldsymbol{V}_C^\perp = \boldsymbol{I}_{n-p}$.

Substituting the SVD of $\boldsymbol{C}$ back into the aforementioned expression of $\mathcal{P}_{\mathcal{C}}$ yields

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \boldsymbol{V}_C^{\perp}(\boldsymbol{V}_C^{\perp})^{\top}\boldsymbol{x} + \tilde{\boldsymbol{d}}, \tag{3.34}$$

where $\tilde{\boldsymbol{d}} = \boldsymbol{V}_C\boldsymbol{\Sigma}_C^{-1}\boldsymbol{U}_C^{\top}\boldsymbol{d} = \boldsymbol{C}^{\dagger}\boldsymbol{d}$.

Step 2: From (3.34), we obtain the difference between the two projections of $\boldsymbol{x}+\boldsymbol{\delta}$ and $\boldsymbol{x}$ onto $\mathcal{C}$, for any $\boldsymbol{x}, \boldsymbol{\delta} \in \mathbb{R}^n$, as $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta}) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \boldsymbol{V}_C^{\perp}(\boldsymbol{V}_C^{\perp})^{\top}\boldsymbol{\delta}$. Using Definition 3.1 with the note that $\Pi_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})$ is always singleton, we have $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at every $\boldsymbol{x} \in \mathbb{R}$ with

$$\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \boldsymbol{V}_C^{\perp}(\boldsymbol{V}_C^{\perp})^{\top},\ c_1(\boldsymbol{x}) = \infty,\ c_2(\boldsymbol{x}) = 0. \tag{3.35}$$

Due to the independence from $\boldsymbol{x}$, we also have $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at every $\boldsymbol{x}^* \in \mathcal{C}$ with

$$\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{V}_C^{\perp}(\boldsymbol{V}_C^{\perp})^{\top},\ c_1(\boldsymbol{x}^*) = \infty,\ c_2(\boldsymbol{x}^*) = 0.$$

Next, substituting $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{V}_C^{\perp}(\boldsymbol{V}_C^{\perp})^{\top}$ into the stationarity equation (3.8) yields $\boldsymbol{V}_C^{\perp}(\boldsymbol{V}_C^{\perp})^{\top}\boldsymbol{A}^{\top}\big(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\big) = \boldsymbol{0}$. Since $\boldsymbol{V}_C^{\perp} \in \mathbb{R}^{n\times(n-p)}$ has full-rank, we can omit the left most $\boldsymbol{V}_C^{\perp}$ and obtain the condition for $\boldsymbol{x}^* \in \mathcal{C}$ to be a Lipschitz stationary point of (3.33) as

$$(\boldsymbol{A}\boldsymbol{V}_C^{\perp})^{\top}(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) = \boldsymbol{0}, \tag{3.36}$$

which means $\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}$ is in the left null space of $\boldsymbol{A}\boldsymbol{V}_C^{\perp}$.[5]

Step 3: Evaluating the projection in (3.34) at $\boldsymbol{z}_\eta^* = \boldsymbol{x}^* - \eta\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$ and using the stationarity condition (3.36) to eliminate the term $\eta\boldsymbol{V}_C^{\perp}(\boldsymbol{V}_C^{\perp})^\top\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$, we have $\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*) = \boldsymbol{x}^*$ for any $\eta > 0$. Thus, the condition in this step for $\boldsymbol{x}^*$ to be a fixed point of Algorithm 3.1 is $\eta > 0$. In addition, substituting $\boldsymbol{x} = \boldsymbol{z}_\eta^*$ into (3.35), we obtain $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at $\boldsymbol{z}_\eta^*$ with

$$\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*) = \boldsymbol{V}_C^{\perp}(\boldsymbol{V}_C^{\perp})^\top,\ c_1(\boldsymbol{z}_\eta^*) = \infty,\ c_2(\boldsymbol{z}_\eta^*) = 0.$$

Step 4: Since $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*) = \nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{V}_C^{\perp}(\boldsymbol{V}_C^{\perp})^\top$, using (3.19), we obtain the asymptotic linear rate as

$$\rho = \max\{|1 - \eta\lambda_1|, |1 - \eta\lambda_{n-p}|\}, \tag{3.37}$$

where λ_1 and λ_{n-p} are the largest and smallest eigenvalues of $(\boldsymbol{A}\boldsymbol{V}_C^{\perp})^\top\boldsymbol{A}\boldsymbol{V}_C^{\perp}$, respectively.

Step 5: From (3.37), we have $\rho < 1$ if and only if $(\boldsymbol{A}\boldsymbol{V}_C^{\perp})^\top\boldsymbol{A}\boldsymbol{V}_C^{\perp}$ has full rank and $0 < \eta < 2/\|\boldsymbol{A}\boldsymbol{V}_C^{\perp}\|_2^2$. It is noted that the latter condition is sufficient for the condition $\eta > 0$ in Step 3.

Step 6: Since $c_1(\boldsymbol{x}^*) = c_1(\boldsymbol{z}_\eta^*) = \infty$ and $c_2(\boldsymbol{x}^*) = c_2(\boldsymbol{z}_\eta^*) = 0$, the region of convergence given by (3.20) is the entire space $\mathbb{R}^n$, which implies global convergence.

Remark 3.4. *The explicit expression of the convergence rate in (3.37) offers a*

[5]Here, it is interesting to note that any stationary point of (3.33) is a global minimizer since (3.33) is a convex optimization problem.

simple method to select the optimal step size:

$$\begin{aligned}\eta_{opt} &= \underset{0<\eta<2/\|\boldsymbol{A}\boldsymbol{V}_C^{\perp}\|_2^2}{\operatorname{argmin}} \max\{|1-\eta\lambda_1|, |1-\eta\lambda_{n-p}|\} \\ &= \frac{2}{\lambda_1 + \lambda_{n-p}}. \end{aligned} \tag{3.38}$$

Using $\eta = \eta_{opt}$, we obtain the optimal rate of convergence

$$\rho_{opt} = 1 - \frac{2}{\kappa((\boldsymbol{A}\boldsymbol{V}_C^{\perp})^{\top}\boldsymbol{A}\boldsymbol{V}_C^{\perp}) + 1}. \tag{3.39}$$

As a comparison, the optimal convergence rate of gradient descent for the unconstrained problem is given by [167]

$$u_{opt} = 1 - \frac{2}{\kappa(\boldsymbol{A}^{\top}\boldsymbol{A}) + 1}.$$

Recall from Remark 3.2 that $\rho_{opt} \leq u_{opt}$ due to the interlacing of eigenvalues of $(\boldsymbol{A}\boldsymbol{V}_C^{\perp})^{\top}\boldsymbol{A}\boldsymbol{V}_C^{\perp}$ and $\boldsymbol{A}^{\top}\boldsymbol{A}$.

3.4.2 Sparse Recovery

In compressed sensing, one would like to reconstruct a sparse signal by finding solutions to under-determined linear systems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ (for $m < n$). This problem can be formulated as an L0-norm constrained

least squares:

$$\boxed{\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2 \quad \text{s.t. } \|\boldsymbol{x}\|_0 \leq s.} \tag{3.40}$$

In the literature, the PGD algorithm for solving (3.40) is often known as iterative hard thresholding (IHT), with myriad applications in medical imaging [56], MIMO communication [75, 195], antenna arrays [196], and scene recognition [234]. The convergence of a special case of IHT in which $\|\boldsymbol{A}\|_2 < 1$ and $\eta = 1$ has been well-studied in [20,21], under the restricted isometry property (RIP) assumption on $\boldsymbol{A}$. In the following, we demonstrate the application of our framework to establishing a local convergence analysis of IHT with a range of different step sizes, without requiring the RIP of $\boldsymbol{A}$.

Step 1: In this example, $\boldsymbol{A}$ and $\boldsymbol{b}$ are given explicitly in (3.40), and the constraint set $\mathcal{C}$ is the closed non-convex set of s-sparse vectors

$$\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_0 \leq s\},$$

with the projection $\mathcal{P}_{\mathcal{C}} : \mathbb{R}^n \to \mathbb{R}^n$ given by [20]

$$[\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})]_i = \begin{cases} 0 & \text{if } |x_i| < |x_{[s]}| \\ x_i & \text{if } |x_i| \geq |x_{[s]}| \end{cases} \text{ for } i = 1, \ldots, n, \tag{3.41}$$

where x_i and $x_{[s]}$ denote the ith coordinate and the sth largest (in magnitude) element of a vector $\boldsymbol{x} \in \mathbb{R}^n$, respectively. In the case $\boldsymbol{x}$ has multiple elements with

the same magnitude as $x_{[s]}$, e.g., $x_{[s]} = x_{[s+1]} > 0$, we sort these entries based on the (descending) lexicographical order so that (3.41) is well-defined (see [20]-p. 10).

Step 2: In contrast to the previous example, the projection here is nonlinear and non-unique since the set $\mathcal{C}$ is a real algebraic variety but not smooth in those points in $\mathbb{R}^n$ of sparsity strictly less than s. The smooth part of $\mathcal{C}$ is the subset

$$\mathcal{C}_{=s} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_0 = s\}$$

of vectors with exactly s non-zero elements. In Appendix 3.6.8, we show that any $\boldsymbol{x}^* \in \Phi_{=s}$ and $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, |x^*_{[s]}|/\sqrt{2})$ share the same index set of s-largest elements (in magnitude), denoted by $\Omega_s(\boldsymbol{x}^*)$.[6] Let the indices in $\Omega_s(\boldsymbol{x}^*)$ be $i_1 \leq \ldots \leq i_s$ and $\boldsymbol{S}_{\boldsymbol{x}^*} = [\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_s}] \in \mathbb{R}^{n\times s}$. Then, we have $(\boldsymbol{S}_{\boldsymbol{x}^*})^\top \boldsymbol{S}_{\boldsymbol{x}^*} = \boldsymbol{I}_s$ and

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}_{\boldsymbol{x}^*}^\top \boldsymbol{x}, \qquad \forall \boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, |x^*_{[s]}|/\sqrt{2}). \tag{3.42}$$

By Definition 3.1, we obtain $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at any $\boldsymbol{x}^* \in \Phi_{=s}$ with

$$\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}_{\boldsymbol{x}^*}^\top,\ c_1(\boldsymbol{x}^*) = \frac{1}{\sqrt{2}}|x^*_{[s]}|,\ c_2(\boldsymbol{x}^*) = 0.$$

[6] It is interesting to note that $|x^*_{[s]}|/\sqrt{2}$ is the largest possible radius. A counter-example is also constructed in Appendix 3.6.8.

Similar to the previous example, the stationarity equation for $\boldsymbol{x}^* \in \mathcal{C}_{=s}$ is given by

$$(\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) = \boldsymbol{0}. \tag{3.43}$$

Thus, we obtain the conditions for $\boldsymbol{x}^* \in \mathcal{C}$ to be a Lipschitz stationary point of (3.40) are $\boldsymbol{x}^* \in \mathcal{C}_{=s}$ and the vector $\boldsymbol{v}^* = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$ satisfies $v_i^* = 0$ for all $i \in \Omega_s(\boldsymbol{x}^*)$.

Step 3: First, following a similar approach to that in [20], we show that the condition in this step for $\boldsymbol{x}^*$ to be a fixed point of Algorithm 3.1 is

$$0 < \eta < \frac{|x^*_{[s]}|}{\|\boldsymbol{v}^*\|_\infty}. \tag{3.44}$$

Since $v_i^* = 0$ for all $i \in \Omega_s(\boldsymbol{x}^*)$, we have $\boldsymbol{z}_\eta^* = \boldsymbol{x}^* - \eta\boldsymbol{v}^*$ satisfies $(\boldsymbol{z}_\eta^*)_i = \boldsymbol{x}_i^*$ for all $i \in \Omega_s(\boldsymbol{x}^*)$. Moreover, for any indices $i \in \Omega_s(\boldsymbol{x}^*)$ and $j \in \{1, \ldots, n\} \setminus \Omega_s(\boldsymbol{x}^*)$, we have

$$\begin{aligned} |(z_\eta^*)_j| &= |x_j^* - \eta v_j^*| = \eta|v_j^*| \\ &< \frac{|x^*_{[s]}|}{\|\boldsymbol{v}^*\|_\infty}|v_j^*| \leq |x^*_{[s]}| \leq |x_i^*| = |(z_\eta^*)_i|, \end{aligned}$$

where the second inequality stems from $|v_j^*| \leq \|\boldsymbol{v}^*\|_\infty$. Therefore, $\Omega_s(\boldsymbol{x}^*)$ contains the s-largest (in magnitude) elements of $\boldsymbol{z}_\eta^*$, and hence, $\boldsymbol{x}^* = \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)$.

Second, we consider the Lipschitz-continuous differentiability of $\mathcal{P}_{\mathcal{C}}$ at $\boldsymbol{z}_\eta^*$. Given η in (3.44), by the same argument as in Appendix 3.6.8, one can show that every point in $\mathcal{B}(\boldsymbol{z}_\eta^*, (|(z_\eta^*)_{[s]}| - |(z_\eta^*)_{[s+1]}|)/\sqrt{2})$ shares the same index set of s-

largest elements (in magnitude) with $\boldsymbol{z}_\eta^*$, which is $\Omega_s(\boldsymbol{x}^*)$. Here, we note that $|(z_\eta^*)_{[s]}| - |(z_\eta^*)_{[s+1]}| = |x_{[s]}^*| - \eta\|\boldsymbol{v}^*\|_\infty$. Thus, we obtain $\mathcal{P}_\mathcal{C}$ is Lipschitz-continuously differentiable at $\boldsymbol{z}_\eta^*$ with

$$\nabla\mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*) = \boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}_{\boldsymbol{x}^*}^\top,$$
$$c_1(\boldsymbol{z}_\eta^*) = \frac{1}{\sqrt{2}}\big(|x_{[s]}^*| - \eta\|\boldsymbol{v}^*\|_\infty\big), \quad c_2(\boldsymbol{z}_\eta^*) = 0.$$

Step 4: Since $\nabla\mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*) = \nabla\mathcal{P}_\mathcal{C}(\boldsymbol{x}^*) = \boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}_{\boldsymbol{x}^*}^\top$, using (3.19), we obtain the asymptotic linear rate as

$$\rho = \max\{|1 - \eta\lambda_1|, |1 - \eta\lambda_s|\}. \tag{3.45}$$

where λ_1 and λ_s are the largest and smallest eigenvalues of $(\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}$, respectively.

Step 5: From (3.45), $\rho < 1$ if and only if $(\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}$ has full rank and

$$0 < \eta < \frac{2}{\|\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}\|_2^2}. \tag{3.46}$$

Combining (3.44) and (3.46) yields the condition on the step size

$$0 < \eta < \min\left\{\frac{2}{\|\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}\|_2^2}, \frac{|x_{[s]}^*|}{\|\boldsymbol{v}^*\|_\infty}\right\}. \tag{3.47}$$

Here, we note that the condition $(\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}$ has full rank is related to the restricted isometry property (RIP) assumption on $\boldsymbol{A}$: $(1 - \delta_s)\|\boldsymbol{x}\|^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|^2 \leq$

$(1+\delta_s)\|\boldsymbol{x}\|^2$, for $\delta_s \in (0,1)$ and any s-sparse vector $\boldsymbol{x} \in \mathbb{R}^n$ [36]. In the reduced-form, we can rewrite the RIP assumption as

$$0 < (1-\delta_s)\|\boldsymbol{y}\|^2 \leq \|\boldsymbol{A}\boldsymbol{S}\boldsymbol{y}\|^2 \leq (1+\delta_s)\|\boldsymbol{y}\|^2, \tag{3.48}$$

for any $\boldsymbol{y} \in \mathbb{R}^s$ and any selection matrix $\boldsymbol{S} \in \mathbb{R}^{n\times s}$ obtained by randomly choosing s columns from the $n \times n$ identity matrix. Substituting $\boldsymbol{S} = \boldsymbol{S}_{\boldsymbol{x}^*}$ into (3.48), we obtain $(\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}$ has full rank.

Step 6: Recall that $c_2(\boldsymbol{x}^*) = c_2(\boldsymbol{z}_\eta^*) = 0$. From (3.20), the region of convergence is given by

$$\|\boldsymbol{x}-\boldsymbol{x}^*\| < \min\left\{\frac{|x^*_{[s]}|}{\sqrt{2}}, \frac{|x^*_{[s]}| - \eta\|\boldsymbol{v}^*\|_\infty}{\sqrt{2}\|\boldsymbol{I}_n - \eta\boldsymbol{A}^\top\boldsymbol{A}\|_2}\right\}. \tag{3.49}$$

Remark 3.5. *Similar to (3.38) and (3.39), the optimal step size and the optimal convergence rate are given by*

$$\begin{aligned}\eta_{opt} &= \frac{2}{\lambda_1((\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}) + \lambda_s((\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})},\\ \rho_{opt} &= 1 - \frac{2}{\kappa((\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}) + 1}.\end{aligned} \tag{3.50}$$

We consider the following numerical experiment to verify the analytical rate in (3.45). We start by generating $\boldsymbol{A}$, $\boldsymbol{x}^*$, *and* $\boldsymbol{b}$ *as follows. First, we sample an* 200×300 *sensing matrix* $\boldsymbol{A}$ *with i.i.d Gaussian distributed entries* $\mathcal{N}(0, 1/200)$.[7]

[7]Note that such random matrix is shown to satisfy the RIP constraint [36].

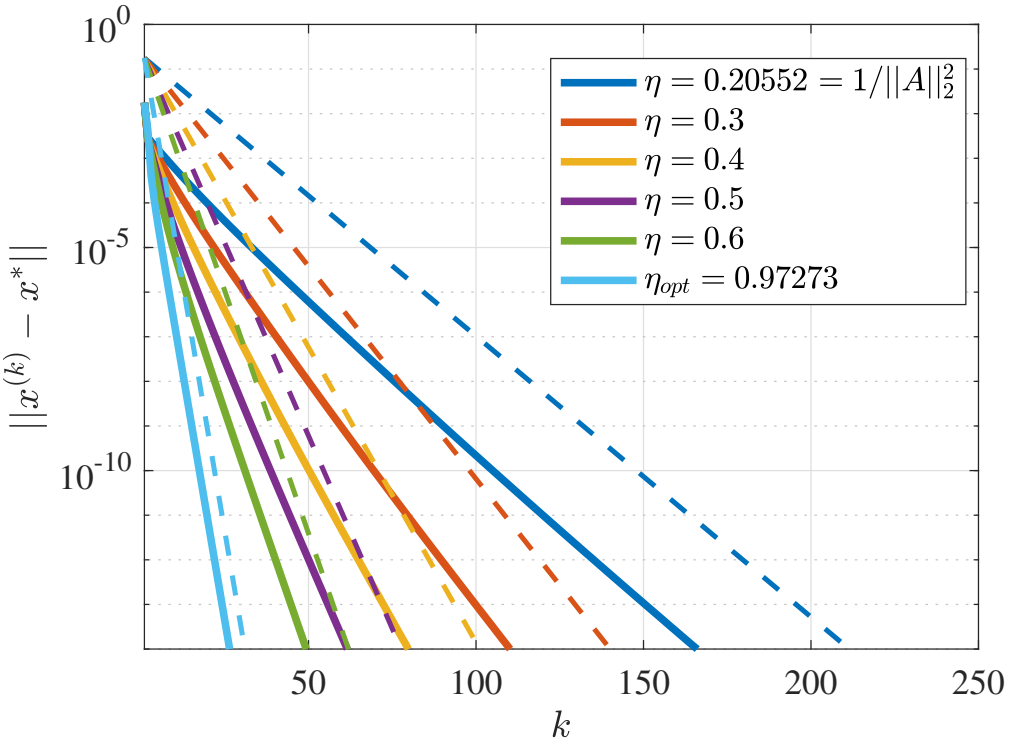


Figure 3.2: (Log-scale) plot of the distance between the current iterate and the local minimizer of the sparse recovery problem, as a function of the number of iterations. Each solid line corresponds to PGD with a different fixed step size. Each dashed line represents the respective exponential bound ρ^k up to a constant, where the theoretical rate ρ is given by (3.45). In the experiment, we select $m = 200, n = 300$, and $s = 10$. The optimal step size $\eta_{opt} = 0.97273$ is computed by (3.50), with the corresponding optimal rate $\rho_{opt} = 0.3613$.

Next, we create a 10*-sparse solution* $\boldsymbol{x}^*$ *by randomly selecting* 10 *coordinates and assigning non-zero values to them based on i.i.d normal distribution* $\mathcal{N}(0,1)$*. Finally, we set* $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^*$*. We apply PGD with different step sizes (listed in Fig. 3.2) including* ρ_{opt} *in (3.50) and record the value of* $\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|$ *as a function of* k*. In Fig. 3.2, the aforementioned curves are presented along with their analytic bounds given by* ρ^k *(up to a constant). The match in the slope between the analytic rate curve and the empirical rate curve verifies the analytic rate predicts accurately the asymptotic rate obtained empirically.*

Remark 3.6. *In Appendix 3.6.8, we further show that any stationary point* $\boldsymbol{x}^*$ *must be a local minimum of (3.40). Moreover, the condition* $(\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}$ *has full rank in Step 5 implies* $\boldsymbol{x}^*$ *is a* ***strict*** *local minimum of (3.40). Finally, it is interesting to note that in [20], the authors assume* $\|\boldsymbol{A}\|_2 < 1$ *and select* $\eta = 1$*. With these assumptions, the rate in (3.45) simplifies to* $\rho = 1 - \lambda_s((\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}) = \left\|\boldsymbol{I}_s - (\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}\right\|_2$*, which is consistent with Eqn. (3.9) in [20].*

3.4.3 Least Squares with the Unit Norm Constraint

A common constraint that arises in regularization methods for ill-posed problems is the spherical constraint [87, 149, 202]. In particular, we consider the following optimization problem

$$\boxed{\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 \quad \text{s.t. } \|\boldsymbol{x}\| = 1,} \tag{3.51}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$.

Step 1: In this example, $\boldsymbol{A}$ and $\boldsymbol{b}$ are given explicitly in (3.51), and the constraint set $\mathcal{C}$ is the closed non-convex sphere

$$\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\},$$

with the projection $\mathcal{P}_{\mathcal{C}} : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \begin{cases} \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} & \text{if } \boldsymbol{x} \neq \boldsymbol{0}, \\ \boldsymbol{e}_1 & \text{if } \boldsymbol{x} = \boldsymbol{0}. \end{cases} \tag{3.52}$$

Step 2: In Example 3.1, we showed that the projection onto the unit sphere is Lipschitz-continuously differentiable at any $\boldsymbol{x} \neq \boldsymbol{0}$. Since $\boldsymbol{0} \notin \mathcal{C}$, we have $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at every $\boldsymbol{x}^* \in \mathcal{C}$ with

$$\nabla P_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{I}_n - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top,\ c_1(\boldsymbol{x}^*) = \infty,\ c_2(\boldsymbol{x}^*) = 2.$$

In addition, substituting $\nabla P_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{I}_n - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top$ into the stationarity equation (3.8) yields $\big(\boldsymbol{I}_n - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) = \boldsymbol{0}$. Equivalently, we have

$$\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) = \gamma \boldsymbol{x}^*, \tag{3.53}$$

where $\gamma = (\boldsymbol{x}^*)^\top \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$ is the Lagrange multiplier at $\boldsymbol{x}^*$ (see Lemma 1 in [218]). Thus, we obtain the condition for $\boldsymbol{x}^* \in \mathcal{C}$ to be a Lipschitz stationary

point of (3.51) is $\boldsymbol{x}^*$ and $\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$ are collinear.

Step 3: First, the necessary condition for $\mathcal{P}_\mathcal{C}$ to be Lipschitz-continuously differentiable at $\boldsymbol{z}_\eta^*$ is $\boldsymbol{z}_\eta^* \in \operatorname{singleton} \Pi_\mathcal{C}$, i.e., $\boldsymbol{z}_\eta^* \neq \boldsymbol{0}$. From (3.53), we have $\boldsymbol{z}_\eta^* = \boldsymbol{x}^* - \eta\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) = (1-\eta\gamma)\boldsymbol{x}^*$. Hence, $\boldsymbol{z}_\eta^* \neq \boldsymbol{0}$ is equivalent to $1-\eta\gamma \neq 0$. Now the projection $\mathcal{P}_\mathcal{C}$ at $\boldsymbol{z}_\eta^* \neq \boldsymbol{0}$ is given by

$$\mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*) = \frac{(1-\eta\gamma)\boldsymbol{x}^*}{\|(1-\eta\gamma)\boldsymbol{x}^*\|} = \frac{1-\eta\gamma}{|1-\eta\gamma|}\boldsymbol{x}^*,$$

which implies $\boldsymbol{x}^* = \mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*)$ if and only if $1-\eta\gamma > 0$. Thus, we obtain the condition for $\eta > 0$ such that $\boldsymbol{x}^*$ is a fixed point of Algorithm 3.1 and $\mathcal{P}_\mathcal{C}$ is Lipschitz-continuously differentiable at $\boldsymbol{z}_\eta^*$ is $1-\eta\gamma > 0$, which is equivalent to

$$\begin{cases} \eta \in (0, \infty) & \text{if } \gamma \leq 0, \\ \eta \in (0, \frac{1}{\gamma}) & \text{if } \gamma > 0. \end{cases} \tag{3.54}$$

Second, it follows from (3.6) that $\mathcal{P}_\mathcal{C}$ is Lipschitz-continuously differentiable at $\boldsymbol{z}_\eta^*$ with

$$\nabla P_\mathcal{C}(\boldsymbol{z}_\eta^*) = \frac{1}{\|\boldsymbol{z}_\eta^*\|}\left(\boldsymbol{I}_n - \frac{\boldsymbol{z}_\eta^*(\boldsymbol{z}_\eta^*)^\top}{\|\boldsymbol{z}_\eta^*\|^2}\right) = \frac{\boldsymbol{I}_n - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top}{1-\eta\gamma},$$
$$c_1(\boldsymbol{z}_\eta^*) = \infty, \quad c_2(\boldsymbol{z}_\eta^*) = \frac{2}{\|\boldsymbol{z}_\eta^*\|^2} = \frac{2}{(1-\eta\gamma)^2}.$$

Step 4: Denote $\boldsymbol{P}_{\boldsymbol{x}^*}^\perp = \boldsymbol{I}_n - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top$. From (3.12), the asymptotic linear rate is

given by

$$\begin{aligned}\rho &= \rho\bigg(\frac{1}{1-\eta\gamma}\boldsymbol{P}_{\boldsymbol{x}^*}^{\perp}(\boldsymbol{I}_n - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{P}_{\boldsymbol{x}^*}^{\perp}\bigg) \\ &= \frac{1}{1-\eta\gamma}\|\boldsymbol{P}_{\boldsymbol{x}^*}^{\perp}(\boldsymbol{I}_n - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{P}_{\boldsymbol{x}^*}^{\perp}\|_2.\end{aligned}$$

Let $\boldsymbol{P}_{\boldsymbol{x}^*}^{\perp} = \boldsymbol{U}_{\boldsymbol{x}^*}\boldsymbol{U}_{\boldsymbol{x}^*}^\top$, where $\boldsymbol{U}_{\boldsymbol{x}^*} \in \mathbb{R}^{n\times(n-1)}$ is a semi-orthogonal matrix whose columns provide a basis for the null space of $\boldsymbol{x}^*$. Then, following the same derivation as in the proof of Corollary 3.1, we obtain

$$\rho = \frac{1}{1-\eta\gamma}\max\{|1-\eta\lambda_1|, |1-\eta\lambda_{n-1}|\}, \tag{3.55}$$

where λ_1 and λ_{n-1} are the largest and smallest eigenvalues of $(\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*})^\top\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}$, respectively.

Step 5: Since $|1-\eta\lambda|/(1-\eta\gamma) < 1$ is equivalent to $\eta\gamma - 1 < 1 - \eta\lambda < 1 - \eta\gamma$, we have $\rho < 1$ if and only if

$$\gamma < \lambda_{n-1} \tag{3.56}$$

and

$$\eta\Big(\gamma + \lambda_1\Big) < 2. \tag{3.57}$$

Similar to (3.54), the inequality in (3.57) can be rewritten as

$$\begin{cases} \eta \in (0, \infty) & \text{if } \gamma \leq -\lambda_1, \\ \eta \in (0, \frac{2}{\gamma + \lambda_1}) & \text{if } \gamma > -\lambda_1. \end{cases}$$

Finally, we note that conditions (3.56) and (3.57) together imply the condition $1 - \eta\gamma > 0$ in Step 3 since $2\eta\gamma < \eta(\gamma + \lambda_{n-1}) \leq \eta(\gamma + \lambda_1) < 2$.

Step 6: To determine the region of linear convergence, we first recall that $c_1(\boldsymbol{x}^*) = c_1(\boldsymbol{z}_\eta^*) = \infty$. Second, we have

$$\|\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)\|_2 = \left\| \frac{\boldsymbol{I}_n - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top}{1 - \eta\gamma} \right\|_2 = \frac{1}{1 - \eta\gamma}.$$

Third, since $\boldsymbol{H} = \boldsymbol{P}_{\boldsymbol{x}^*}^{\perp}(\boldsymbol{I}_n - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{P}_{\boldsymbol{x}^*}^{\perp}/(1 - \eta\gamma)$ is symmetric, one can choose $\boldsymbol{Q}$ in the eigendecomposition $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$ to be orthogonal, with $\kappa(\boldsymbol{Q}) = 1$. Thus, from (3.13), we obtain the region of linear convergence as

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \frac{1 - \rho}{2(t^2 + t)}, \tag{3.58}$$

where $t = \|\boldsymbol{I}_n - \eta \boldsymbol{A}^\top \boldsymbol{A}\|_2/(1 - \eta\gamma)$.

Remark 3.7. *The local linear rate in (3.55) matches the rate provided by Theorem 1 in [218]. Compared to the setting in [218], here we consider a special case of the quadratic that is convex (and hence, $\lambda_d \geq 0$). By minimizing the rate in (3.55) over η, we also obtain the same optimal rate of linear convergence given by*

Lemma 5 in [218]:

$$\rho_{opt} = \frac{\lambda_1 - \lambda_{n-1}}{\lambda_1 + \lambda_{n-1} - 2\gamma} \text{ with } \eta_{opt} = \frac{2}{\lambda_1 + \lambda_{n-1}}.$$

Interestingly, condition (3.56) implies $\boldsymbol{x}^*$ is a strict local minimum of (3.51) (see Lemma 2 in [218]). Since $\rho < 1$ is one of the conditions in Theorem 3.1, our analysis requires $\boldsymbol{x}^*$ to be a strict local minimum of (3.51) in order to obtain linear convergence. Finally, our framework provides the region of linear convergence in (3.58), which is not given in [218].

3.4.4 Matrix Completion

3.4.4.1 Background

The last application is an application of our framework to the matrix case. In matrix completion [33], given a rank-r matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ (for $1 \leq r \leq \min\{m, n\}$) with a set of its observed entries indexed by Ω, of cardinality $0 < s < mn$, we wish to recover the unknown entries of $\boldsymbol{M}$ in the complement set $\bar{\Omega}$ by solving the following optimization:

$$\boxed{\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} \frac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{X} - \boldsymbol{M})\|_F^2 \text{ s.t. } \operatorname{rank}(\boldsymbol{X}) \leq r,} \tag{3.59}$$

where $\mathcal{P}_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is the orthogonal projection onto the set of $m \times n$ matrices supported in Ω, i.e.,

$$[\mathcal{P}_\Omega(\boldsymbol{X})]_{ij} = \begin{cases} X_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{if } (i,j) \notin \Omega. \end{cases}$$

It is noted that while $\boldsymbol{M}$ is unknown, the projection $\mathcal{P}_\Omega(\boldsymbol{M})$ is unambiguously determined by the observed entries in $\boldsymbol{M}$. In the literature, the PGD algorithm for solving (3.59) is also known as the Singular Value Projection (SVP) algorithm for matrix completion [43, 55, 104, 105], with the update

$$\boldsymbol{X}^{(k+1)} = \mathcal{P}_{\mathcal{M}_{\le r}}\big(\boldsymbol{X}^{(k)} - \eta\mathcal{P}_\Omega(\boldsymbol{X}^{(k)} - \boldsymbol{M})\big).$$

Here, $\mathcal{M}_{\le r}$ is the set of matrices of rank at most r, i.e.,

$$\mathcal{M}_{\le r} = \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \operatorname{rank}(\boldsymbol{X}) \le r\}.$$

In addition, the orthogonal projection $\mathcal{P}_{\mathcal{M}_{\le r}} : \mathbb{R}^{m\times n} \to \mathcal{M}_{\le r}$ is defined by Eckart–Young–Mirsky theorem [61] as follows. Let $\operatorname{SVD}(\boldsymbol{X})$ be the set of all triples $(\boldsymbol{\Sigma}, \boldsymbol{U}, \boldsymbol{V})$ such that $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and

$$\begin{cases} \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1(\boldsymbol{X}), \ldots, \sigma_n(\boldsymbol{X})), \\ \boldsymbol{U} \in \mathbb{R}^{m\times n}, \boldsymbol{V} \in \mathbb{R}^{n\times n} : \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_n. \end{cases}$$

Denote $\boldsymbol{u}_i(\boldsymbol{X})$ and $\boldsymbol{v}_i(\boldsymbol{X})$ the ith columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Then, the set of all projections of $\boldsymbol{X}$ onto $\mathcal{M}_{\leq r}$ is given by

$$\Pi_{\mathcal{M}_{\leq r}}(\boldsymbol{X}) = \Big\{ \sum_{i=1}^{r} \sigma_i(\boldsymbol{X}) \boldsymbol{u}_i(\boldsymbol{X}) \boldsymbol{v}_i(\boldsymbol{X})^\top \mid (\boldsymbol{\Sigma}, \boldsymbol{U}, \boldsymbol{V}) \in \mathrm{SVD}(\boldsymbol{X}) \Big\}. \tag{3.60}$$

The set $\Pi_{\mathcal{M}_{\leq r}}(\boldsymbol{X})$ is singleton if and only if $\sigma_r(\boldsymbol{X}) = 0$ or $\sigma_r(\boldsymbol{X}) > \sigma_{r+1}(\boldsymbol{X})$. In the case $\Pi_{\mathcal{M}_{\leq r}}(\boldsymbol{X})$ has multiple elements, we define $\mathcal{P}_{\mathcal{M}_{\leq r}}(\boldsymbol{X})$ as the greatest element in $\Pi_{\mathcal{M}_{\leq r}}(\boldsymbol{X})$ based on the lexicographical order. We re-emphasize that our subsequent analysis holds independently of this choice.

In differential geometry, it is well-known that $\mathcal{M}_{\leq r}$ is a closed set of $\mathbb{R}^{m \times n}$ but non-smooth in those points of rank strictly less than r [126]. Similar to sparse recovery, the smooth part of $\mathcal{M}_{\leq r}$ is the set of matrices of fixed rank r:

$$\mathcal{M}_{=r} = \{ \boldsymbol{X} \in \mathbb{R}^{m \times n} \mid \mathrm{rank}(\boldsymbol{X}) = r \}.$$

At any $\boldsymbol{X}^* \in \mathcal{M}_{=r}$, it is shown [211] that derivative of $\mathcal{P}_{\mathcal{M}_{\leq r}}$ is a linear mapping from $\mathbb{R}^{m \times n}$ to $\mathbb{R}^{m \times n}$ satisfying

$$\nabla \mathcal{P}_{\mathcal{M}_{\leq r}}(\boldsymbol{X}^*)(\boldsymbol{\Delta}) = \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_\perp}, \tag{3.61}$$

where $\boldsymbol{P}_{\boldsymbol{U}_\perp}$ and $\boldsymbol{P}_{\boldsymbol{V}_\perp}$ are the projections onto the left and right null spaces of $\boldsymbol{X}^*$, respectively. More importantly, for any $\boldsymbol{\Delta} \in \mathbb{R}^{m \times n}$, Theorem 3 in [211] asserts

that

$$\sup_{\boldsymbol{Y}\in\Pi_{\mathcal{M}_{\leq r}}(\boldsymbol{X}^*+\boldsymbol{\Delta})}\|\boldsymbol{Y}-\boldsymbol{X}^*-\nabla\mathcal{P}_{\mathcal{M}_{\leq r}}(\boldsymbol{X}^*)(\boldsymbol{\Delta})\|_F\leq\frac{4(1+\sqrt{2})}{\sigma_r(\boldsymbol{X}^*)}\|\boldsymbol{\Delta}\|_F^2. \tag{3.62}$$

3.4.4.2 Vectorized version of matrix completion

To apply our proposed framework to matrix completion, we consider a vectorized version of (3.59) as follows. Slightly extending the notation, we denote $\mathcal{C}=\{\mathrm{vec}(\boldsymbol{X})\mid\boldsymbol{X}\in\mathcal{M}_{\leq r}\}$ and $\mathrm{vec}(\Omega)=\{(j-1)m+i\mid(i,j)\in\Omega\}$ with s distinct elements $1\leq i_1<\ldots<i_s\leq mn$. Let $\boldsymbol{S}_\Omega=[\boldsymbol{e}_{i_1},\ldots,\boldsymbol{e}_{i_s}]\in\mathbb{R}^{mn\times s}$ be the selection matrix satisfying

$$\begin{cases}\boldsymbol{S}_\Omega^\top\boldsymbol{S}_\Omega=\boldsymbol{I}_s,\\ \mathrm{vec}\big(\mathcal{P}_\Omega(\boldsymbol{X})\big)=\boldsymbol{S}_\Omega\boldsymbol{S}_\Omega^\top\mathrm{vec}(\boldsymbol{X}).\end{cases}$$

Then, problem (3.59) can be represented as

$$\min_{\boldsymbol{x}\in\mathbb{R}^{mn}}\frac{1}{2}\|\boldsymbol{S}_\Omega\boldsymbol{S}_\Omega^\top\boldsymbol{x}-\boldsymbol{S}_\Omega\boldsymbol{S}_\Omega^\top\mathrm{vec}(\boldsymbol{M})\|^2\ \text{s.t.}\ \boldsymbol{x}\in\mathcal{C}.$$

Step 1: In this vectorized version of matrix completion, we have $\boldsymbol{A}=\boldsymbol{S}_\Omega\boldsymbol{S}_\Omega^\top$, $\boldsymbol{b}=\boldsymbol{S}_\Omega\boldsymbol{S}_\Omega^\top\mathrm{vec}(\boldsymbol{M})$, and $\mathcal{C}$ is a closed non-convex set. For any vector $\boldsymbol{x}\in\mathbb{R}^{mn}$, let $\boldsymbol{X}=\mathrm{vec}^{-1}(\boldsymbol{x})$ with $\mathcal{P}_{\mathcal{M}_{\leq r}}(\boldsymbol{X})=\sum_{i=1}^r\sigma_i(\boldsymbol{X})\boldsymbol{u}_i(\boldsymbol{X})\boldsymbol{v}_i(\boldsymbol{X})^\top$, for some $(\boldsymbol{\Sigma},\boldsymbol{U},\boldsymbol{V})\in\mathrm{SVD}(\boldsymbol{X})$. The projection $\mathcal{P}_{\mathcal{C}}$ is given by $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})=\mathrm{vec}(\sum_{i=1}^r\sigma_i(\boldsymbol{X})\boldsymbol{u}_i(\boldsymbol{X})\boldsymbol{v}_i(\boldsymbol{X})^\top)$. Using the fact that $\mathrm{vec}(\boldsymbol{u}\boldsymbol{v}^\top)=\boldsymbol{v}\otimes\boldsymbol{u}$, for any vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ of compatible

dimensions, $\mathcal{P}_{\mathcal{C}}$ can then be represented as

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \sum_{i=1}^{r} \sigma_i(\boldsymbol{X})\big(\boldsymbol{v}_i(\boldsymbol{X}) \otimes \boldsymbol{u}_i(\boldsymbol{X})\big). \tag{3.63}$$

Step 2: In the following, we show that $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at any point in the set

$$\mathcal{C}_{=r} = \{\mathrm{vec}(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{M}_{=r}\}.$$

In particular, for any $\boldsymbol{x}^* \in \mathcal{C}_{=r}$, we prove that $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at $\boldsymbol{x}^*$ with

$$\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{P}^{\perp}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp}, \quad c_1(\boldsymbol{x}^*) = \infty, \quad c_2(\boldsymbol{x}^*) = \frac{4(1+\sqrt{2})}{\sigma_r(\boldsymbol{X}^*)},$$

where $\boldsymbol{X}^* = \mathrm{vec}^{-1}(\boldsymbol{x}^*)$. Indeed, the constants $c_1(\boldsymbol{x}^*)$ and $c_2(\boldsymbol{x}^*)$ are obtained from the matrix inequality form (3.62). Regarding $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)$, let $\boldsymbol{P}_{\boldsymbol{U}_\perp}$ and $\boldsymbol{P}_{\boldsymbol{V}_\perp}$ be the projections onto the left and right null spaces of $\boldsymbol{X}^*$, respectively. Denote $\boldsymbol{P}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp} = \boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}$ and $\boldsymbol{P}^{\perp}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp} = \boldsymbol{I}_{mn} - \boldsymbol{P}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp}$. Since $\mathrm{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^\top \otimes \boldsymbol{A})\,\mathrm{vec}(\boldsymbol{B})$, for any matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ of compatible dimensions, (3.61) can be vectorized to obtain $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)(\boldsymbol{\delta}) = (\boldsymbol{I}_{mn} - \boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})\boldsymbol{\delta} = \boldsymbol{P}^{\perp}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp}\boldsymbol{\delta}$ for any $\boldsymbol{\delta} \in \mathbb{R}^{mn}$.

Next, the stationarity condition (3.8) can be represented using $\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)(\boldsymbol{\delta}) = \boldsymbol{P}^{\perp}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp}\boldsymbol{\delta}$ as $\boldsymbol{P}^{\perp}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp}\boldsymbol{S}_\Omega\boldsymbol{S}_\Omega^\top\big(\boldsymbol{S}_\Omega\boldsymbol{S}_\Omega^\top\boldsymbol{x}^* - \boldsymbol{S}_\Omega\boldsymbol{S}_\Omega^\top\mathrm{vec}(\boldsymbol{M})\big) = \boldsymbol{0}$. Denote $\boldsymbol{Q}_\perp \in \mathbb{R}^{mn \times r(m+n-r)}$ the matrix satisfying $\boldsymbol{Q}_\perp^\top\boldsymbol{Q}_\perp = \boldsymbol{I}_{r(m+n-r)}$ and $\boldsymbol{Q}_\perp\boldsymbol{Q}_\perp^\top = \boldsymbol{P}^{\perp}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp}$. Then, we obtain

the conditions for $\boldsymbol{x}^*$ to be a Lipschitz stationary point of (3.59) are $\boldsymbol{x}^* \in \mathcal{C}_{=r}$ and

$$\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \big(\boldsymbol{x}^* - \mathrm{vec}(\boldsymbol{M})\big) = \boldsymbol{0}. \tag{3.64}$$

Step 3: The stationarity condition (3.64) leads to two cases. The first case is when $\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega$ has full (row-)rank and hence,

$$\boldsymbol{S}_\Omega^\top \big(\boldsymbol{x}^* - \mathrm{vec}(\boldsymbol{M})\big) = \boldsymbol{0}. \tag{3.65}$$

In matrix form, (3.65) can be rewritten as $\mathcal{P}_\Omega(\boldsymbol{X}^*) = \mathcal{P}_\Omega(\boldsymbol{M})$, which implies $\boldsymbol{X}^*$ is a global minimizer of (3.59). Interestingly, this case enjoys the special setting considered in Corollary 3.1 as

$$\boldsymbol{z}_\eta^* = \boldsymbol{x}^* - \eta \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \big(\boldsymbol{x}^* - \mathrm{vec}(\boldsymbol{M})\big) = \boldsymbol{x}^*, \tag{3.66}$$

for any $\eta > 0$. In the second case, if $\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega$ has rank strictly less than $r(m+n-r)$, then $\boldsymbol{S}_\Omega^\top \big(\boldsymbol{x}^* - \mathrm{vec}(\boldsymbol{M})\big)$ may not be $\boldsymbol{0}$ (e.g., a non-zero right singular vector of $\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega$). This implies $\boldsymbol{z}_\eta^* \neq \boldsymbol{x}^*$ and one needs to characterize the Lipschitz-continuous differentiability of the projection $\mathcal{P}_\mathcal{C}$ onto the set of low-rank matrices at $\boldsymbol{z}_\eta^*$ that may not have exact rank r. While the derivative of $\mathcal{P}_\mathcal{C}$ at a matrix with rank greater than r has been studied in [64,211], it requires complete development of the error bound on the first-order expansion of this operator to obtain the constants $c_1(\boldsymbol{z}_\eta^*)$ and $c_2(\boldsymbol{z}_\eta^*)$. For the purpose of demonstration, we restrict our subsequent analysis to the first case when $\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega$ has full (row-)rank. Since $\boldsymbol{z}_\eta^* = \boldsymbol{x}^*$

in this case, $\mathcal{P}_{\mathcal{C}}$ is Lipschitz-continuously differentiable at $\boldsymbol{z}_\eta^*$ with

$$\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*) = \boldsymbol{P}^{\perp}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp},\ c_1(\boldsymbol{z}_\eta^*) = \infty,\ c_2(\boldsymbol{z}_\eta^*) = \frac{4(1+\sqrt{2})}{\sigma_r(\boldsymbol{X}^*)}.$$

Step 4: Since $\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*) = \nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{P}^{\perp}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp}$, using (3.19), we obtain the asymptotic linear rate as

$$\rho = \max\{|1-\eta\lambda_1|, |1-\eta\lambda_{r(m+n-r)}|\}, \tag{3.67}$$

where λ_1 and $\lambda_{r(m+n-r)}$ are the largest and smallest eigenvalues of $\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp$, respectively.

Step 5: From (3.67), we have $\rho < 1$ if and only if $\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp$ has full rank and

$$0 < \eta < \frac{2}{\|\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp\|_2}.$$

Here, we would like to point out the condition $\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp$ has full rank implies $s \geq r(m+n-r)$, which can be interpreted as a requirement for the number of observations being no less than the degree of freedom in matrix completion. The invertibility of $\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp$ is also equivalent to the injectivity of the sampling operator restricted to the tangent space T to $\mathcal{M}_{\leq r}$ at $\boldsymbol{X}^*$, denoted by $\mathcal{A}_{\Omega T}$ in [33]-Section 4.2. It is interesting to note that under the standard assumptions on uniform sampling and incoherence property, Candès and Recht [33] showed that $\mathcal{A}_{\Omega T}$ is injective with high probability.

Step 6: Recall that $c_1(\boldsymbol{x}^*) = c_1(\boldsymbol{z}_\eta^*) = \infty$. Since $\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*) = \nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{P}^{\perp}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp}$,

using (3.20), the region of linear convergence is given by

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \frac{(1-\rho)\sigma_r(\boldsymbol{X}^*)}{8(1+\sqrt{2})}. \tag{3.68}$$

Remark 3.8. *Similar to (3.38) and (3.39), the optimal step size and the optimal convergence rate are given by*

$$\begin{aligned} \eta_{opt} &= \frac{2}{\lambda_1(\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp) + \lambda_{r(m+n-r)}(\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp)}, \\ \rho_{opt} &= 1 - \frac{2}{\kappa(\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp) + 1}. \end{aligned} \tag{3.69}$$

We consider the following numerical experiment to verify the analytical rate in (3.67). The data is generated randomly as follows. First, we sample two matrices $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *with i.i.d normally distributed entries, of dimensions* 50×3 *and* 40×3*, respectively. Next, we obtain the rank-*3 *matrix of dimension* 50×40 *as the product* $\boldsymbol{X}^* = \boldsymbol{A}\boldsymbol{B}^\top$*. Third, we select* 800 *observations uniformly at random among the* 2000 *positions in* $\boldsymbol{X}^*$*. We apply PGD with different step sizes (listed in Fig. 3.3) including* η_{opt} *in (3.69) and record the value of* $\left\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^*\right\|_F$ *as a function of* k*. It can be seen from Fig. 3.3 that the theoretical rate matches well the empirical rate, reassuring the correctness of our analysis in the previous section.*

Remark 3.9. *The rate in (3.67) has not been proposed in the literature. However, in the special case of using unit step size, it matches the rate established for the IHTSVD algorithm in [46]. In their work, the authors provide the result relative to the matrix* $(\boldsymbol{S}_{\bar{\Omega}})^\top \boldsymbol{P}_{\boldsymbol{U}_\perp \boldsymbol{V}_\perp} \boldsymbol{S}_{\bar{\Omega}}$ *instead of* $\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp$*, where* $\boldsymbol{S}_{\bar{\Omega}} \in \mathbb{R}^{mn \times (mn-s)}$ *is the*

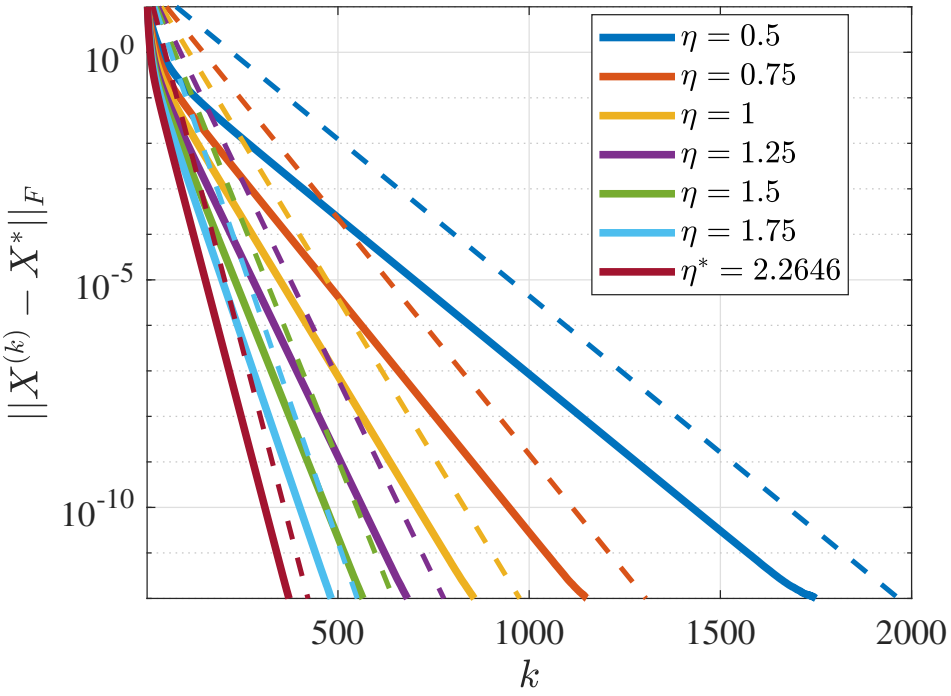


Figure 3.3: (Log-scale) plot of the distance between the current iterate and the local minimizer of the matrix completion problem, as a function of the number of iterations. Each solid line corresponds to PGD with a different fixed step size. Each dashed line represents the respective exponential bound ρ^k up to a constant, where the theoretical rate ρ is given by (3.67). In the experiment, we select $m = 50, n = 40, r = 3$, and $s = 800$. The optimal step size $\eta_{opt} = 2.2833$ is given by (3.69), with the corresponding optimal rate $\rho_{opt} = 0.9265$.

selection matrix that is complement to $\boldsymbol{S}_\Omega$*. It can be shown that the two matrices share the same set of eigenvalues while may only differ by the eigenvalues at* 1*. Since IHTSVD uses* $\eta = 1$*, these unit eigenvalues do not affect the maximization in (3.67). Compared to the local convergence result in [46], our application in this subsection not only considers PGD with different step sizes but also includes the region of linear convergence in (3.68).*

3.5 Conclusion and Future Work

We presented a unified framework to analyze the local convergence of projected gradient descent for constrained least squares. Our analysis provides the asymptotic rate of convergence in a closed-form expression, the number of iterations required to reach certain accuracy, and the local region of convergence. Notably, our technique relies on the Lipschitz-continuous differentiability of the projection operator at two key points: $\boldsymbol{x}^*$ and $\boldsymbol{z}_\eta^*$. Finally, we demonstrated the application of our proposed framework to local convergence analysis of PGD in four well-known problems: linear equality-constrained least squares, sparse recovery, least squares with a unit norm constraint, and matrix completion.

While the work here focuses on the specific setting of linear converges of the PGD algorithm, we believe it can be expanded in several directions. First, our framework can be utilized to analyze the following cases: *(i)* adaptive step size schemes (e.g., the backtracking line search rule), *(ii)* accelerated methods (e.g., the Nesterov's accelerated gradient and the Heavy Ball method), *(iii)* general ob-

jective functions other than least squares, and *(iv)* other algorithms for manifold optimization such as Riemannian gradient descent. Another interesting research direction is to sharpen the theoretical bound on the ROC in order to better explain the actual region in which the algorithm converges to the desired solution. Finally, the proposed framework can be used to further study the performance of PGD for a variety of constrained least squares problems arising in the area of phase-only beamforming [206], online power system optimization [94], spectral compressed sensing [28], and linear dimensionality reduction [50].

3.6 Appendix

3.6.1 Proof of Lemma 3.1

Our goal in the proof of Lemma 3.1 is to show that if the fixed point condition $\boldsymbol{x}^* = \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* - \eta \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}))$ holds, then the stationarity condition $\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)\boldsymbol{A}^\top\big(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\big) = \boldsymbol{0}$ holds. Note that if $\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) = \boldsymbol{0}$, then the stationarity condition holds trivially. Hence, we focus on the proof for $\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) \neq \boldsymbol{0}$. We first show that for any $0 \leq \alpha < 1$, $\boldsymbol{x}^* = \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* - \eta \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}))$ is a sufficient condition for

$$\Pi_{\mathcal{C}}\big(\boldsymbol{x}^* - \alpha\eta \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})\big) = \{\boldsymbol{x}^*\}. \tag{3.70}$$

Then, using (3.70) and the differentiability of $\mathcal{P}_{\mathcal{C}}$ at $\boldsymbol{x}^*$, we prove $\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)\boldsymbol{A}^\top\big(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\big) = \boldsymbol{0}$. We proceed with the detailed proof.

First, let $\boldsymbol{v}^* = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$ and $\boldsymbol{z}^*_{\alpha\eta} = \boldsymbol{x}^* - \alpha\eta\boldsymbol{v}^*$. On the one hand, for any

$0 \leq \alpha < 1$ and $\boldsymbol{y} \in \Pi_{\mathcal{C}}(\boldsymbol{z}^*_{\alpha\eta})$, we have

$$\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{z}^*_\eta\| &= \|(\boldsymbol{y} - \boldsymbol{z}^*_{\alpha\eta}) + (\boldsymbol{z}^*_{\alpha\eta} - \boldsymbol{z}^*_\eta)\| \\
&\leq \|\boldsymbol{y} - \boldsymbol{z}^*_{\alpha\eta}\| + \|\boldsymbol{z}^*_{\alpha\eta} - \boldsymbol{z}^*_\eta\| \\
&= d(\boldsymbol{z}^*_{\alpha\eta}, \mathcal{C}) + \|\boldsymbol{z}^*_{\alpha\eta} - \boldsymbol{z}^*_\eta\| \\
&\leq \|\boldsymbol{x}^* - \boldsymbol{z}^*_{\alpha\eta}\| + \|\boldsymbol{z}^*_{\alpha\eta} - \boldsymbol{z}^*_\eta\|, \qquad (3.71)
\end{aligned}$$

where the first inequality uses the triangle inequality that holds when $\boldsymbol{y} - \boldsymbol{z}^*_{\alpha\eta} = \beta(\boldsymbol{z}^*_{\alpha\eta} - \boldsymbol{z}^*_\eta)$, for some $\beta \geq 0$. Using the fact that $\boldsymbol{z}^*_\eta = \boldsymbol{x}^* - \eta\boldsymbol{v}^*$ and $\boldsymbol{z}^*_{\alpha\eta} = \boldsymbol{x}^* - \alpha\eta\boldsymbol{v}^*$, we obtain

$$\begin{aligned}
\|\boldsymbol{x}^* - \boldsymbol{z}^*_{\alpha\eta}\| + \|\boldsymbol{z}^*_{\alpha\eta} - \boldsymbol{z}^*_\eta\| &= \|\alpha\eta\boldsymbol{v}^*\| + \|(1-\alpha)\eta\boldsymbol{v}^*\| \\
&= \|\alpha\eta\boldsymbol{v}^* + (1-\alpha)\eta\boldsymbol{v}^*\| \\
&= \|\eta\boldsymbol{v}^*\| \\
&= \|\boldsymbol{x}^* - \boldsymbol{z}^*_\eta\|. \qquad (3.72)
\end{aligned}$$

From (3.71) and (3.72), we have

$$\|\boldsymbol{y} - \boldsymbol{z}^*_\eta\| \leq \|\boldsymbol{x}^* - \boldsymbol{z}^*_\eta\|, \qquad (3.73)$$

with the equality holding if and only if

$$\begin{cases} \boldsymbol{y} - \boldsymbol{z}^*_{\alpha\eta} = \beta(\boldsymbol{z}^*_{\alpha\eta} - \boldsymbol{z}^*_{\eta}), \\ \|\boldsymbol{y} - \boldsymbol{z}^*_{\alpha\eta}\| = \|\boldsymbol{x}^* - \boldsymbol{z}^*_{\alpha\eta}\|. \end{cases} \tag{3.74}$$

Using the fact that $\boldsymbol{z}^*_{\eta} = \boldsymbol{x}^* - \eta\boldsymbol{v}^*$ and $\boldsymbol{z}^*_{\alpha\eta} = \boldsymbol{x}^* - \alpha\eta\boldsymbol{v}^*$, (3.74) holds if and only if

$$\beta = \frac{\alpha}{1-\alpha} \text{ and } \boldsymbol{y} = \boldsymbol{x}^*. \tag{3.75}$$

On the other hand, since $\boldsymbol{x}^* = \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}^*_{\eta})$ and $\boldsymbol{y} \in \mathcal{C}$, we have

$$\|\boldsymbol{y} - \boldsymbol{z}^*_{\eta}\| \geq \|\boldsymbol{x}^* - \boldsymbol{z}^*_{\eta}\|. \tag{3.76}$$

From (3.73) and (3.76), we conclude that $\|\boldsymbol{y} - \boldsymbol{z}^*_{\eta}\| = \|\boldsymbol{x}^* - \boldsymbol{z}^*_{\eta}\|$. Moreover, from (3.75), the equality holds if and only if $\boldsymbol{y} = \boldsymbol{x}^*$. Since this holds for any $\boldsymbol{y} \in \Pi_{\mathcal{C}}(\boldsymbol{z}^*_{\alpha\eta})$, we conclude that $\Pi_{\mathcal{C}}(\boldsymbol{z}^*_{\alpha\eta}) = \{\boldsymbol{x}^*\}$ for all $0 \leq \alpha < 1$.

Next, using the differentiability of the projection $\mathcal{P}_{\mathcal{C}}$ at $\boldsymbol{x}^*$ from Definition 3.1, we have

$$\lim_{\alpha\to 0} \sup_{\boldsymbol{y}\in\Pi_{\mathcal{C}}(\boldsymbol{x}^*-\alpha\eta\boldsymbol{v}^*)} \frac{\|\boldsymbol{y} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) - \nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)(\alpha\eta\boldsymbol{v}^*)\|}{\|\alpha\eta\boldsymbol{v}^*\|} = 0.$$

Substituting $\Pi_{\mathcal{C}}(\boldsymbol{x}^* - \alpha\eta\boldsymbol{v}^*) = \Pi_{\mathcal{C}}(\boldsymbol{z}^*_{\alpha\eta}) = \{\boldsymbol{x}^*\}$ and $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{x}^*$ into the last

equation, we obtain

$$\begin{aligned} 0 &= \lim_{\alpha \to 0} \frac{\|\boldsymbol{x}^* - \boldsymbol{x}^* - \alpha\eta \nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)\boldsymbol{v}^*\|}{\|\alpha\eta\boldsymbol{v}^*\|} \\ &= \lim_{\alpha \to 0} \frac{\alpha\eta \, \|\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)\boldsymbol{v}^*\|}{\alpha\eta \, \|\boldsymbol{v}^*\|} \\ &= \frac{\|\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)\boldsymbol{v}^*\|}{\|\boldsymbol{v}^*\|}, \end{aligned}$$

which only holds if $\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)\boldsymbol{v}^* = \boldsymbol{0}$. This completes our proof of the lemma.

3.6.2 Proof of Corollary 3.1

In the following, under the assumption $\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*) = \nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*) = \boldsymbol{U}_{\boldsymbol{x}^*}\boldsymbol{U}_{\boldsymbol{x}^*}^\top$, we show that *(i)* the asymptotic convergence rate $\rho(\boldsymbol{H})$ is given by (3.19), *(ii)* the sufficient conditions for $\rho(\boldsymbol{H}) < 1$ are $(\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}$ is full rank and (3.18) holds, and *(iii)* the region of linear convergence can be simplified from (3.13) to (3.20).

First, we prove (3.19) by simplifying the expression of $\boldsymbol{H}$ in (3.17) and the fact that $\boldsymbol{U}_{\boldsymbol{x}^*}^\top \boldsymbol{U}_{\boldsymbol{x}^*} = \boldsymbol{I}_d$ as follows. Substituting $\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)$ and $\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)$ by $\boldsymbol{U}_{\boldsymbol{x}^*}\boldsymbol{U}_{\boldsymbol{x}^*}^\top$ into (3.12) yields

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{U}_{\boldsymbol{x}^*}\boldsymbol{U}_{\boldsymbol{x}^*}^\top(\boldsymbol{I}_n - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{U}_{\boldsymbol{x}^*}\boldsymbol{U}_{\boldsymbol{x}^*}^\top \\ &= \boldsymbol{U}_{\boldsymbol{x}^*}(\boldsymbol{I}_d - \eta \boldsymbol{U}_{\boldsymbol{x}^*}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*})\boldsymbol{U}_{\boldsymbol{x}^*}^\top, \end{aligned}$$

where the second equality stems from $\boldsymbol{U}_{\boldsymbol{x}^*}^\top \boldsymbol{U}_{\boldsymbol{x}^*} = \boldsymbol{I}_d$. Since $\boldsymbol{H}$ is symmetric, its

spectral radius equals to its spectral norm:

$$\rho(\boldsymbol{H}) = \|\boldsymbol{U}_{\boldsymbol{x}^*}(\boldsymbol{I}_d - \eta\boldsymbol{U}_{\boldsymbol{x}^*}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*})\boldsymbol{U}_{\boldsymbol{x}^*}^\top\|_2.$$

Using the fact that the spectral norm is invariant under left-multiplication by matrices with orthonormal columns and right-multiplication by matrices with orthonormal rows (see [151] - Exercise 5.6.9), we further have

$$\rho(\boldsymbol{H}) = \|\boldsymbol{I}_d - \eta\boldsymbol{U}_{\boldsymbol{x}^*}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}\|_2. \tag{3.77}$$

Let $\boldsymbol{U}_{\boldsymbol{x}^*}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{U}}^\top$ be an eigendecomposition, where $\hat{\boldsymbol{U}} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix and $\hat{\boldsymbol{\Lambda}} = \mathrm{diag}(\lambda_1(\boldsymbol{U}_{\boldsymbol{x}^*}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}), \ldots, \lambda_d(\boldsymbol{U}_{\boldsymbol{x}^*}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}))$. Since $\hat{\boldsymbol{U}}^\top\hat{\boldsymbol{U}} = \boldsymbol{I}_d$, (3.77) can be represented as

$$\rho(\boldsymbol{H}) = \left\|\hat{\boldsymbol{U}}(\boldsymbol{I}_d - \eta\hat{\boldsymbol{\Lambda}})\hat{\boldsymbol{U}}^\top\right\|_2 = \left\|\boldsymbol{I}_d - \eta\hat{\boldsymbol{\Lambda}}\right\|_2.$$

Now using the fact that the spectral norm of a diagonal matrix is the maximum of the absolute values of its diagonal entries, we obtain

$$\begin{aligned}\rho(\boldsymbol{H}) &= \max_{1\le i\le d}|1 - \eta\lambda_i(\boldsymbol{U}_{\boldsymbol{x}^*}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*})| \\ &= \max\{|1 - \eta\lambda_1|, |1 - \eta\lambda_d|\}.\end{aligned}$$

Second, we establish the sufficient conditions for $\rho(\boldsymbol{H}) < 1$ by bounding each

term inside the maximum in (3.19) as follows. Since $\lambda_d \geq 0$, we have

$$-1 < 1 - \eta\lambda_1 \leq 1 - \eta\lambda_d \leq 1, \text{ for } i = 1, \ldots, d,$$

if $0 < \eta < 2/\lambda_1$. It is also noted from the definition of the spectral norm that $\|\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}\|_2^2 = \lambda_1$. Therefore, $\rho(\boldsymbol{H}) \leq 1$ provided that (3.18) holds. The equality $\rho(\boldsymbol{H}) = 1$ holds if and only if $\lambda_d = 0$, i.e., $(\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}$ is singular. In other words, when $(\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*})^\top \boldsymbol{A}\boldsymbol{U}_{\boldsymbol{x}^*}$ is full rank and (3.18) holds, the linear convergence is guaranteed as $\rho(\boldsymbol{H}) < 1$.

Finally, the region of linear convergence in (3.20) is determined based on simplifying (3.13) as follows. First, using Remark 3.1, we obtain $\kappa(\boldsymbol{Q}) = 1$. Second, from (3.17), we have $\|\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)\|_2 = \|\mathcal{P}_{T_{\boldsymbol{x}^*}(\mathcal{C})}\|_2 = 1$. Third, substituting $\kappa(\boldsymbol{Q}) = 1$ and $\|\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)\|_2 = 1$ into (3.13) yields (3.20). This completes our proof of the corollary.

3.6.3 Proof of Lemma 3.2

Our goal is to show the error vector $\boldsymbol{\delta}^{(k)}$ satisfies the asymptotically-linear quadratic system dynamic in (3.25) and to bound the norm of the residual $\boldsymbol{q}_2$ by (3.26).

First, our key idea in proving (3.25) is the Lipschitz-continuous differentiability of $\mathcal{P}_{\mathcal{C}}$ at $\boldsymbol{x}^*$ and at $\boldsymbol{z}_\eta^*$. Specifically, for any k such that $\boldsymbol{\delta}^{(k)}$ admits a perturbation

$(\boldsymbol{I} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}$ that satisfies

$$\|(\boldsymbol{I} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}\| < c_1(\boldsymbol{z}_\eta^*), \tag{3.78}$$

applying the Lipschitz-continuous differentiability of $\mathcal{P}_\mathcal{C}$ at $\boldsymbol{z}_\eta^*$ to (3.23) yields

$$\boldsymbol{\delta}^{(k+1)} = \nabla \mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*)(\boldsymbol{I} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)} + \boldsymbol{q}_1(\boldsymbol{\delta}^{(k)}), \tag{3.79}$$

where the residual $\boldsymbol{q}_1 : \mathbb{R}^n \to \mathbb{R}^n$ satisfies

$$\begin{aligned} \|\boldsymbol{q}_1(\boldsymbol{\delta}^{(k)})\| &\leq c_2(\boldsymbol{z}_\eta^*)\|(\boldsymbol{I} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}\|^2 \\ &\leq c_2(\boldsymbol{z}_\eta^*)u_\eta^2\|\boldsymbol{\delta}^{(k)}\|^2. \end{aligned} \tag{3.80}$$

On the other hand, using the fact that $\boldsymbol{x}^* = \mathcal{P}_\mathcal{C}(\boldsymbol{x}^*)$, $\boldsymbol{x}^{(k)} = \mathcal{P}_\mathcal{C}(\boldsymbol{x}^{(k)})$, and the Lipschitz-continuous differentiability of $\mathcal{P}_\mathcal{C}$ at $\boldsymbol{x}^*$ with the perturbation $\boldsymbol{\delta}^{(k)} \in \mathcal{B}(\boldsymbol{0}, c_1(\boldsymbol{x}^*))$, we obtain

$$\begin{aligned} \boldsymbol{\delta}^{(k)} &= \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \\ &= \mathcal{P}_\mathcal{C}(\boldsymbol{x}^* + \boldsymbol{\delta}^{(k)}) - \mathcal{P}_\mathcal{C}(\boldsymbol{x}^*) \\ &= \nabla \mathcal{P}_\mathcal{C}(\boldsymbol{x}^*)\boldsymbol{\delta}^{(k)} + \boldsymbol{q}_{\boldsymbol{x}^*}(\boldsymbol{\delta}^{(k)}), \end{aligned} \tag{3.81}$$

where the residual $\boldsymbol{q}_{\boldsymbol{x}^*} : \mathbb{R}^n \to \mathbb{R}^n$ satisfies $\|\boldsymbol{q}_{\boldsymbol{x}^*}(\boldsymbol{\delta}^{(k)})\| \leq c_2(\boldsymbol{x}^*)\|\boldsymbol{\delta}^{(k)}\|^2$. We proceed with the proof of (3.25) by combining the results from (3.79) and (3.81) as follows. Since $\|(\boldsymbol{I} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}\| \leq \|\boldsymbol{I} - \eta \boldsymbol{A}^\top \boldsymbol{A}\|_2\|\boldsymbol{\delta}^{(k)}\| = u_\eta\|\boldsymbol{\delta}^{(k)}\|$, the sufficient condition

for (3.78) is $\|\boldsymbol{\delta}^{(k)}\| < c_1(\boldsymbol{z}_\eta^*)/u_\eta$. Thus, $\|\boldsymbol{\delta}^{(k)}\| < \min\{c_1(\boldsymbol{x}^*), c_1(\boldsymbol{z}_\eta^*)/u_\eta\}$ is sufficient for both (3.79) and (3.81). Substituting (3.81) into the RHS of (3.79), we obtain (3.25) with $\boldsymbol{q}_2(\boldsymbol{\delta}^{(k)}) = \nabla\mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*)(\boldsymbol{I} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{q}_{\boldsymbol{x}^*}(\boldsymbol{\delta}^{(k)}) + \boldsymbol{q}_1(\boldsymbol{\delta}^{(k)})$. Next, to bound the norm of the residual $\boldsymbol{q}_2$, we apply the triangle inequality as follows

$$\|\boldsymbol{q}_2(\boldsymbol{\delta}^{(k)})\| \leq \|\nabla\mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*)(\boldsymbol{I} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{q}_{\boldsymbol{x}^*}(\boldsymbol{\delta}^{(k)})\| + \|\boldsymbol{q}_1(\boldsymbol{\delta}^{(k)})\|. \tag{3.82}$$

On the one hand, the first term on the RHS of (3.82) can be bounded by

$$\begin{aligned}\|\nabla\mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*)(\boldsymbol{I} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{q}_{\boldsymbol{x}^*}(\boldsymbol{\delta}^{(k)})\| &\leq \|\nabla\mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*)\|_2\|\boldsymbol{I} - \eta\boldsymbol{A}^\top\boldsymbol{A}\|_2\|\boldsymbol{q}_{\boldsymbol{x}^*}(\boldsymbol{\delta}^{(k)})\| \\ &\leq \|\nabla\mathcal{P}_\mathcal{C}(\boldsymbol{z}_\eta^*)\|_2 u_\eta c_2(\boldsymbol{x}^*)\|\boldsymbol{\delta}^{(k)}\|^2. \end{aligned} \tag{3.83}$$

On the other hand, the second term on the RHS of (3.82) can be bounded by (3.80). Combining the two bounds, we obtain (3.26).

3.6.4 Proof of Lemma 3.3

In this section, our goal is to show the recursion on the transformed error vector (3.29) holds at any $k \in \mathbb{N}$ provided that the initial error vector lies within the region of linear convergence described by (3.13). In the first step, we prove that if the current transformed error vector lies within the region of linear convergence

$$\|\tilde{\boldsymbol{\delta}}^{(k)}\| < \min\Big\{\frac{c_1(\boldsymbol{x}^*)}{\|\boldsymbol{Q}\|_2}, \frac{c_1(\boldsymbol{z}_\eta^*)}{\|\boldsymbol{Q}\|_2 u_\eta}, \frac{1-\rho(\boldsymbol{H})}{q/\|\boldsymbol{Q}^{-1}\|_2}\Big\}. \tag{3.84}$$

then $\tilde{\boldsymbol{\delta}}^{(k+1)} = \boldsymbol{\Lambda}\tilde{\boldsymbol{\delta}}^{(k)} + \boldsymbol{q}_3(\tilde{\boldsymbol{\delta}}^{(k)})$ and moreover, the next transformed error vector also lies within the region of linear convergence

$$\|\tilde{\boldsymbol{\delta}}^{(k+1)}\| < \min\Big\{\frac{c_1(\boldsymbol{x}^*)}{\|\boldsymbol{Q}\|_2}, \frac{c_1(\boldsymbol{z}_\eta^*)}{\|\boldsymbol{Q}\|_2 u_\eta}, \frac{1-\rho(\boldsymbol{H})}{q/\|\boldsymbol{Q}^{-1}\|_2}\Big\}. \tag{3.85}$$

Therefore, by the principle of induction, the initial condition on the transformed error vector, i.e., (3.84) holds at $k = 0$, is the sufficient condition for (3.29) to hold at any $k \in \mathbb{N}$. In the second step, we show that (3.84) holds at $k = 0$ if the initial condition on the error vector (3.13) holds and hence, completes the proof of lemma. We proceed with our detailed proof below.

First, let us assume that (3.84) holds. We have

$$\begin{aligned}
\|\boldsymbol{\delta}^{(k)}\| = \|\boldsymbol{Q}\tilde{\boldsymbol{\delta}}^{(k)}\| &\leq \|\boldsymbol{Q}\|_2\|\tilde{\boldsymbol{\delta}}^{(k)}\| \\
&< \|\boldsymbol{Q}\|_2 \min\Big\{\frac{c_1(\boldsymbol{x}^*)}{\|\boldsymbol{Q}\|_2}, \frac{c_1(\boldsymbol{z}_\eta^*)}{\|\boldsymbol{Q}\|_2 u_\eta}, \frac{1-\rho(\boldsymbol{H})}{q/\|\boldsymbol{Q}^{-1}\|_2}\Big\} \\
&\leq \min\Big\{c_1(\boldsymbol{x}^*), \frac{c_1(\boldsymbol{z}_\eta^*)}{u_\eta}\Big\},
\end{aligned} \tag{3.86}$$

Thus, by Lemma 3.2, we have $\boldsymbol{\delta}^{(k+1)} = \boldsymbol{H}\boldsymbol{\delta}^{(k)} + \boldsymbol{q}_2(\boldsymbol{\delta}^{(k)})$. Substituting $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$ and multiplying both sides with $\boldsymbol{Q}^{-1}$ yields $\boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(k+1)} = \boldsymbol{\Lambda}\boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(k)} + \boldsymbol{Q}^{-1}\boldsymbol{q}_2(\boldsymbol{\delta}^{(k)})$. Replacing $\boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(k)}$ by $\tilde{\boldsymbol{\delta}}^{(k)}$ and $\boldsymbol{\delta}^{(k)}$ by $\boldsymbol{Q}\tilde{\boldsymbol{\delta}}^{(k)}$ in the last equation, we obtain (3.29),

i.e., $\tilde{\boldsymbol{\delta}}^{(k+1)} = \boldsymbol{\Lambda}\tilde{\boldsymbol{\delta}}^{(k)} + \boldsymbol{q}_3(\tilde{\boldsymbol{\delta}}^{(k)})$. Here, the second term $\boldsymbol{q}_3$ can be bounded as follows

$$\begin{aligned}
\|\boldsymbol{q}_3\big(\tilde{\boldsymbol{\delta}}^{(k)}\big)\| &= \|\boldsymbol{Q}^{-1}\boldsymbol{q}_2(\boldsymbol{Q}\tilde{\boldsymbol{\delta}}^{(k)})\| \leq \|\boldsymbol{Q}^{-1}\|_2\|\boldsymbol{q}_2(\boldsymbol{Q}\tilde{\boldsymbol{\delta}}^{(k)})\| \\
&\leq \|\boldsymbol{Q}^{-1}\|_2\big(c_2(\boldsymbol{x}^*) + \|\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)\|_2 c_2(\boldsymbol{z}_\eta^*)\big)\|\boldsymbol{Q}\tilde{\boldsymbol{\delta}}^{(k)}\|^2 \\
&\leq \|\boldsymbol{Q}^{-1}\|_2\big(c_2(\boldsymbol{x}^*) + \|\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{z}_\eta^*)\|_2 c_2(\boldsymbol{z}_\eta^*)\big)\|\boldsymbol{Q}\|_2^2\|\tilde{\boldsymbol{\delta}}^{(k)}\|^2 \\
&= \frac{q}{\|\boldsymbol{Q}^{-1}\|_2}\|\tilde{\boldsymbol{\delta}}^{(k)}\|^2. \qquad (3.87)
\end{aligned}$$

Now, taking the norms of both sides of (3.29) and applying the triangle inequality yield

$$\begin{aligned}
\|\tilde{\boldsymbol{\delta}}^{(k+1)}\| &\leq \|\boldsymbol{\Lambda}\tilde{\boldsymbol{\delta}}^{(k)}\| + \|\boldsymbol{q}_3\big(\tilde{\boldsymbol{\delta}}^{(k)}\big)\| \\
&\leq \rho(\boldsymbol{H})\|\tilde{\boldsymbol{\delta}}^{(k)}\| + \frac{q}{\|\boldsymbol{Q}^{-1}\|_2}\|\tilde{\boldsymbol{\delta}}^{(k)}\|^2 \\
&< \rho(\boldsymbol{H})\|\tilde{\boldsymbol{\delta}}^{(k)}\| + \big(1-\rho(\boldsymbol{H})\big)\|\tilde{\boldsymbol{\delta}}^{(k)}\| \\
&= \|\tilde{\boldsymbol{\delta}}^{(k)}\|, \qquad (3.88)
\end{aligned}$$

where the second inequality stems from $\|\tilde{\boldsymbol{\delta}}^{(k)}\| < (1-\rho(\boldsymbol{H}))/(q/\|\boldsymbol{Q}^{-1}\|_2)$. From (3.84) and (3.88), we conclude that (3.85) holds. By the principle of induction, we have (3.84) holds for all $k \in \mathbb{N}$ provided that it holds at $k = 0$, i.e.,

$$\|\tilde{\boldsymbol{\delta}}^{(0)}\| < \min\Big\{\frac{c_1(\boldsymbol{x}^*)}{\|\boldsymbol{Q}\|_2}, \frac{c_1(\boldsymbol{z}_\eta^*)}{\|\boldsymbol{Q}\|_2 u_\eta}, \frac{1-\rho(\boldsymbol{H})}{q/\|\boldsymbol{Q}^{-1}\|_2}\Big\}. \qquad (3.89)$$

Second, we prove that (3.13) is sufficient for (3.89). Using the definition $\tilde{\boldsymbol{\delta}}^{(k)} =$

$\boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(k)}$, we have

$$\|\tilde{\boldsymbol{\delta}}^{(k)}\| = \|\boldsymbol{Q}^{-1}\boldsymbol{\delta}^{(k)}\| \leq \|\boldsymbol{Q}^{-1}\|_2\|\boldsymbol{\delta}^{(k)}\|. \tag{3.90}$$

Upper-bounding $\|\boldsymbol{\delta}^{(k)}\|$ by the LHS of (3.13) and substituting back into (3.90) yield

$$\|\tilde{\boldsymbol{\delta}}^{(k)}\| < \|\boldsymbol{Q}^{-1}\|_2 \min\Big\{\frac{c_1(\boldsymbol{x}^*)}{\kappa(\boldsymbol{Q})}, \frac{c_1(\boldsymbol{z}_\eta^*)}{\kappa(\boldsymbol{Q})u_\eta}, \frac{1-\rho(\boldsymbol{H})}{q}\Big\}.$$

Finally, replacing $\kappa(\boldsymbol{Q})$ by the product $\|\boldsymbol{Q}\|_2\|\boldsymbol{Q}^{-1}\|_2$ and simplifying yield (3.89). This completes our proof of the lemma.

3.6.5 Proof of Lemma 3.4

In this section, we show the convergence of $\{\|\tilde{\boldsymbol{\delta}}^{(k)}\|\}_{k=0}^{\infty}$ using Theorem 1 in [216]. Our idea is to consider a surrogate sequence $\{a_k\}_{k=0}^{\infty}$ that upper-bounds $\{\|\tilde{\boldsymbol{\delta}}^{(k)}\|\}_{k=0}^{\infty}$:

$$\begin{cases} a_0 = \|\tilde{\boldsymbol{\delta}}^{(0)}\|, \\ a_{k+1} = \rho(\boldsymbol{H})a_k + \frac{q}{\|\boldsymbol{Q}^{-1}\|_2}a_k^2. \end{cases}$$

First, we prove by induction that

$$\|\tilde{\boldsymbol{\delta}}^{(k)}\| \leq a_k \quad \forall k \in \mathbb{N}. \tag{3.91}$$

The base case when $k = 0$ holds trivially as $a_0 = \|\tilde{\boldsymbol{\delta}}^{(0)}\|$. In the induction step, given $\|\tilde{\boldsymbol{\delta}}^{(k)}\| \leq a_k$ for some integer $k \geq 0$, we have

$$\begin{aligned}\|\tilde{\boldsymbol{\delta}}^{(k+1)}\| &\leq \rho(\boldsymbol{H})\|\tilde{\boldsymbol{\delta}}^{(k)}\| + \frac{q}{\|\boldsymbol{Q}^{-1}\|_2}\|\tilde{\boldsymbol{\delta}}^{(k)}\|^2 \\ &\leq \rho a_k + \frac{q}{\|\boldsymbol{Q}^{-1}\|_2}a_k^2 \\ &= a_{k+1}.\end{aligned}$$

By the principle of induction, (3.91) holds for all $k \in \mathbb{N}$. Next, applying Theorem 1 in [216], under the condition $a_0 = \|\tilde{\boldsymbol{\delta}}^{(0)}\| < (1-\rho(\boldsymbol{H}))/(q/\|\boldsymbol{Q}^{-1}\|_2)$, yields $a_k \leq \tilde{\epsilon}a_0$ for any integer k satisfies (3.31). From (3.91), we further have $\|\tilde{\boldsymbol{\delta}}^{(k)}\| \leq a_k \leq \tilde{\epsilon}a_0 = \tilde{\epsilon}\|\tilde{\boldsymbol{\delta}}^{(0)}\|$. This completes our proof of the lemma.

3.6.6 Related Work

In this section, we review existing approaches to convergence analysis of iterative first-order methods in optimization including projected gradient descent. We present several aspects of convergence, namely, convergence to a global versus a local optimum and speed of convergence. Finally, we clarify our contribution in this work with regard to previous works in the literature.

3.6.6.1 Convergence of Iterative First-Order Methods

Convergence properties of iterative algorithms such as PGD often involve two key aspects: the quality of convergent points and the speed of convergence. On the one hand, the quality of convergent points provides useful insights into when the algorithm converges, whether it converges to a stationary point or a set of stationary points of the problem, and how big is the gap between the objective function at the convergent point and the optimal objective value. On the other hand, the speed of convergence concerns the order of convergence, the rate of convergence, and the number of iterations required to obtain sufficiently small errors. Let $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ be the sequence of updates generated by a certain iterative first-order method (e.g., PGD). In order to prove the convergence of the algorithm, it is common [16, 24, 140, 160, 167] to consider the convergence of the following quantities to $\mathbf{0}$ as $k \to \infty$: *(i)* the norm of the generalized gradient ($\|\frac{1}{\eta}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)})\|$), *(ii)* the gap between current objective function and the optimal value ($|f(\boldsymbol{x}^{(k)}) - f^*|$), and *(iii)* the distance to a convergent point ($\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|$). Here, we note that f^* and $\boldsymbol{x}^*$ are the limiting points of the objective function $f(\boldsymbol{x}^{(k)})$ and the parameter $\boldsymbol{x}^{(k)}$ as the number of iterations k goes to infinity, respectively. In (i), the convergence of the generalized gradient norm to 0 implies the stationarity condition of the constrained problem is satisfied. It follows that the algorithm converges to *a set of stationary points* of the problem. In (ii), the convergence on the function side is often obtained via the monotonicity of the objective-value sequence $\{f(\boldsymbol{x}^{(k)})\}_{k=0}^{\infty}$ (e.g., decreasing to a limiting value f^*). This in turn implies the sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$

converges to *a set of local optima* that yields the same objective function value f^*.[8] In (iii), the convergence of $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|$ implies convergence to a unique point that is often an *isolated local optimum point* of the problem. Typically, convergence on the domain side is used in linear convergence proofs for strongly convex settings.

3.6.6.2 Convergence to a Global Optimum

In general, a stationary point can be a saddle point, a local/global minimum, or local/global maximum of the problem. When both the objective function and the constraint set are convex, it is well-known that all stationary points are also global optima of the problem. Convergence analysis of iterative algorithm (e.g., PGD) in convex optimization therefore focus on providing a universal upper bound on the distance to the global solutions. Analysis on the domain side (iii) is usually used in the presence of *strong convexity* that guarantees the *uniqueness* of the global optimum [140]-Section 8.6. Without the strong convexity, one may resort to analysis on the function side (ii) in order to prove convergence to *a set* of global optima [12]-Section 10.4.3. When convexity is not guaranteed, due to a non-convex objective and/or a non-convex constraint set, convergence analysis has recourse to a set of stationary points by bounding the generalized gradient norm through iterations (i) [16]-Section 2.3.2. Notwithstanding, recent advances

[8] An example for such scenario is minimizing a convex but not strongly convex function $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ subject to $\boldsymbol{x} \in \mathbb{R}^n$ and $\|\boldsymbol{x}\|_2^2 = 1$. The $2n$ vectors $\{\boldsymbol{e}_i\}_{i=1}^n$ and $\{-\boldsymbol{e}_i\}_{i=1}^n$ are local minimizers that obtain the same objective function value. It is worthwhile mentioning that they are also the global solutions of the foregoing problem.

in structured non-convex optimization have shed light on convergence guarantees to global solutions of the problem. By exploiting the special structure of some classes of non-convex problems and using appropriate initialization, PGD can be shown to converge to a unique global optimum despite the non-convexity of these problems. Examples of such powerful results include sparse recovery with restricted isometry properties [21], matrix completion with incoherence properties [144], empirical risk minimization with restricted strong convexity and smoothness properties [116], and spherically constrained quadratic minimization with hidden convexity [13].

3.6.6.3 Convergence to a Local Optimum

In general non-convex settings, domain-side convergence analysis is restricted to the local region around the convergence point $\boldsymbol{x}^*$. Such points can be a saddle point, a local minimum, or a local maximum of the problem. The ROC associated with $\boldsymbol{x}^*$ is the neighborhood in which the algorithm (e.g., PGD) is guaranteed to converge to $\boldsymbol{x}^*$ when initialized inside this region. To a certain extent, the ROC in the aforementioned global convergence analysis is the entire feasible space. However, while global convergence analysis does not require the initialization to be close to the global solution, it often ignores the local structure near the solution needed for establishing sharp bounds on the speed of convergence. In particular, bounding techniques employed in global convergence analysis hold universally, including worst-case scenarios. Thus, in many problem-specific settings where the solution lies in a benign neighborhood, the global analysis could lead to conser-

vative convergence rate bounds. As an illustration, in minimizing a smooth and strongly convex function f, gradient descent with a fixed step size achieves the rate of convergence at most $(\kappa - 1)/(\kappa + 1)$ [168], where κ is the (global) condition number of f. Recall that the condition number of a differentiable convex function is the ratio of its smoothness L to strong convexity μ [160]. For any quadratic function, this global bound is also an exact and attainable estimate thanks to the fact that the objective curvature is unchanged everywhere. For non-quadratic objectives, on the other hand, this global bound may be loose as κ takes into account the worst-case scenario, in which the objective function is most ill-conditioned. The asymptotic behavior of gradient descent near the solution indeed relies on the condition number of the local Hessian $\kappa(\boldsymbol{x}^*)$ of the objective function, defining as $\lambda_{\max}(\nabla^2 f(\boldsymbol{x}^*))/\lambda_{\min}(\nabla^2 f(\boldsymbol{x}^*))$. Generally, we have $\mu \leq \lambda_{\min}(\nabla^2 f(\boldsymbol{x})) \leq \lambda_{\max}(\nabla^2 f(\boldsymbol{x}^*)) \leq L$, for any $\boldsymbol{x}$ in the domain of f, which implies $\kappa(\boldsymbol{x}^*) \leq \kappa$. This local condition number $\kappa(\boldsymbol{x}^*)$ can be significantly smaller than the global condition number κ and hence, *a local convergence analysis can yield a tighter bound that reflects the actual convergence speed of the algorithm near the solution*. Similar situation also occurs for constrained least squares in which the Hessian restricted to the constrained set can depend on the local structure of the set.

3.6.6.4 Speed of Convergence

To illustrate the concept of convergence speed, let us consider the convergence on the domain side, i.e., the distance $\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|$. Let μ be a number between 0 and 1. The convergence of $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ to $\boldsymbol{x}^*$ is said to be at rate $\mu \triangleq \mu(\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty})$ if $\mu = \inf_{\{\epsilon_k\}_{k=0}^{\infty}} \lim_{k\to\infty} \epsilon_{k+1}/\epsilon_k$, for any monotonically decreasing sequence $\{\epsilon_k\}_{k=0}^{\infty}$ satisfying $\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\| \leq \epsilon_k$ for all index k. The asymptotic rate of convergence of gradient descent to $\boldsymbol{x}^*$, denoted by ρ, is defined by the worst-case rate of convergence among all possible sequences $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ that are generated by the algorithm and converge to $\boldsymbol{x}^*$, i.e., $\rho = \sup_{\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}} \mu(\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty})$. Depending on the value of ρ in the interval $[0, 1]$, the convergence is said to be *sublinear* when $\rho = 1$, *linear* when $0 < \rho < 1$, or *superlinear* when $\rho = 0$. The lower the value of ρ is, the faster the speed of convergence is and the fewer the number of iterations needed is to obtain a close approximation of the solution. Thus, analytical estimation of the convergence rate plays a pivotal role in convergence analysis. We would like to note two distinct methods for linear convergence rate analysis dating back to the 1960s. The first approach was proposed by Polyak [167], based on his earlier study into nonlinear difference equations [166]. The author analyzed the asymptotic convergence of gradient descent for minimizing some objective function f. Assuming $\boldsymbol{x}^*$ is a non-singular local minimum of f, Polyak showed that for any $\delta > 0$, there exists $\epsilon > 0$ such that if $\left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\| < \epsilon$ then the sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ generated by

gradient descent satisfies

$$\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\| \leq \left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\| (\rho + \delta)^k, \tag{3.92}$$

where $\rho = \max\{|1 - \eta\lambda_{\max}|, |1 - \eta\lambda_{\min}|\}$ and $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $\nabla^2 f(\boldsymbol{x}^*)$, respectively. Here we emphasize that f does not need to be smooth and strongly convex everywhere but only so around $\boldsymbol{x}^*$. By setting $\eta_{opt} = 2/(\lambda_{\max} + \lambda_{\min})$, the optimal rate of convergence is given by $\rho_{opt} = (\kappa^* - 1)/(\kappa^* + 1)$, where $\kappa^* = \lambda_{\max}/\lambda_{\min}$ is the condition number of the local Hessian $\nabla^2 f(\boldsymbol{x}^*)$. When f is a strongly convex quadratic, the local result coincides with the aforementioned global result in [168] ($\kappa^* = \kappa$). The expression of ρ in (3.92) is called the **asymptotic convergence rate** of gradient descent with fixed step size η.[9] The second approach was developed by Daniel [51] in 1967, while studying gradient descent with *exact line search*, i.e., choosing η that minimizes the objective at each iteration. Utilizing the Kantorovich inequality [114], the author proved that if $\boldsymbol{x}^{(0)}$ is sufficiently close to $\boldsymbol{x}^*$, there exist a constant ϵ and a sequence $\{q_k\}_{k=0}^{\infty}$ such that

$$\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\| \leq \epsilon \prod_{i=0}^{k} q_i, \qquad \lim_{k\to\infty} q_k = (\kappa^* - 1)/(\kappa^* + 1).$$

[9] It is worthwhile to mention that using a similar technique, Nesterov [160] proved that the asymptotic rate is at most $\hat{\rho} = (\kappa^* + 1)/(\kappa^* + 3)$. While this bound also exploits the local information of the optimization problem, we note that it is not as tight as the bound in (3.92).

Note that here the characteristics of convergence are also exploited through the Hessian $\nabla^2 f(\boldsymbol{x}^*)$. This result was then extended to study the asymptotic convergence of projected gradient descent for constrained optimization [71, 132, 139].

3.6.7 Proof of Example 3.1

Our goal in this proof is to establish the Lipschitz differentiability of the projection operator onto the unit sphere $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\}$. We start by establishing the Lipschitz differentiability at a point on $\mathcal{C}$ and then extend it to any nonzero point in $\mathbb{R}^n$. For the Lipschitz differentiability on $\mathcal{C}$, we introduce the following lemma:

Lemma 3.5. *For any $\boldsymbol{x}^* \in \mathcal{C}$, we have*

$$\sup_{\boldsymbol{y} \in \Pi_{\mathcal{C}}(\boldsymbol{x}^* + \boldsymbol{\delta})} \|\boldsymbol{y} - \boldsymbol{x}^* - \big(\boldsymbol{I} - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{\delta}\| \leq 2\|\boldsymbol{\delta}\|^2. \tag{3.93}$$

Proof. We consider two cases:

Case 1: If $\boldsymbol{x}^* + \boldsymbol{\delta} = \boldsymbol{0}$, then $\Pi_{\mathcal{C}}(\boldsymbol{0}) = \mathcal{C}$ and $\|\boldsymbol{\delta}\| = \|\boldsymbol{x}^*\| = 1$. For any $\boldsymbol{y} \in \mathcal{C}$, substituting $\boldsymbol{\delta} = -\boldsymbol{x}^*$ and then using the fact that $\boldsymbol{I} - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top$ is the projection onto the null space of $\boldsymbol{x}^*$, we have

$$\begin{aligned} \boldsymbol{y} - \boldsymbol{x}^* - \big(\boldsymbol{I} - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{\delta} &= \boldsymbol{y} - \boldsymbol{x}^* + \big(\boldsymbol{I} - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{x}^* \\ &= \boldsymbol{y} - \boldsymbol{x}^*. \end{aligned}$$

Next, taking the norm and using the triangle inequality yield

$$\begin{aligned}\|\boldsymbol{y}-\boldsymbol{x}^*-\big(\boldsymbol{I}-\boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{\delta}\| &= \|\boldsymbol{y}-\boldsymbol{x}^*\| \\ &\leq \|\boldsymbol{y}\|+\|\boldsymbol{x}^*\| = 2\|\boldsymbol{\delta}\|^2,\end{aligned}$$

where the last step stems from $\|\boldsymbol{y}\| = \|\boldsymbol{x}^*\| = \|\boldsymbol{\delta}\| = 1$. Thus, (3.93) holds in this case.

Case 2: If $\boldsymbol{x}^*+\boldsymbol{\delta} \neq \boldsymbol{0}$, then $\Pi_{\mathcal{C}}(\boldsymbol{x}^*+\boldsymbol{\delta})$ is singleton containing the unique projection

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*+\boldsymbol{\delta}) = \frac{\boldsymbol{x}^*+\boldsymbol{\delta}}{\|\boldsymbol{x}^*+\boldsymbol{\delta}\|}.$$

Hence, (3.93) is equivalent to

$$\left\| \frac{\boldsymbol{x}^*+\boldsymbol{\delta}}{\|\boldsymbol{x}^*+\boldsymbol{\delta}\|} - \boldsymbol{x}^* - \big(\boldsymbol{I}-\boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{\delta} \right\| \leq 2\|\boldsymbol{\delta}\|^2. \tag{3.94}$$

We prove (3.94) by *(i)* showing that for any scalars $u > 0$ and $(1-u)^2 \leq v \leq (1+u)^2$:

$$(17u-2)v^2 - 2u(1-u)^2 v + (1-u)^4(u+2) \geq 0, \tag{3.95}$$

and *(ii)* showing that (3.95) is equivalent to (3.94) with $u = \|\boldsymbol{x}^*+\boldsymbol{\delta}\| > 0$ and $v = \|\boldsymbol{\delta}\|^2 \geq 0$.

(i) To prove (3.95), let us consider the following cases:

1. If $0 < u \leq 2/17$, then for $v \leq (1+u)^2$, we have

$$\begin{aligned}(17u-2)v^2 - 2u(1-u)^2 v + (1-u)^4(u+2) \\ &\geq (17u-2)(1+u)^4 - 2u(1-u)^2(1+u)^2 + (1-u)^4(u+2) \\ &= 16u^2(u+2)(u^2+2u+2) \geq 0.\end{aligned}$$

2. If $2/17 < u \leq 1/2$, then for $(1-u)^2 \leq v \leq (1+u)^2$, the following holds

$$\begin{aligned}(17u-2)v^2 - 2u(1-u)^2 v + (1-u)^4(u+2) \\ &\geq (17u-2)(1-u)^4 - 2u(1-u)^2(1+u)^2 + (1-u)^4(u+2) \\ &= 8u(1-u)^2(2-u)(1-2u) \geq 0.\end{aligned}$$

3. If $u > 1/2$, using the quadratic vertex at $v = u(1-u)^2/(17u-2)$ as the minimum point, we obtain

$$(17u-2)v^2 - 2u(1-u)^2 v + (1-u)^4(u+2) \geq \frac{4(1-u)^4(4u^2+8u-1)}{17u-2} \geq 0.$$

(ii) Now for $u = \|\boldsymbol{x}^* + \boldsymbol{\delta}\| > 0$ and $v = \|\boldsymbol{\delta}\|^2 \geq 0$, we have $(\boldsymbol{x}^*)^\top \boldsymbol{\delta} = (u^2 - v - 1)/2$

and

$$
\begin{aligned}
(3.94) \Leftrightarrow \quad & \left\| \frac{\boldsymbol{x}^* + \boldsymbol{\delta}}{\|\boldsymbol{x}^* + \boldsymbol{\delta}\|} - \boldsymbol{x}^* - \big(\boldsymbol{I} - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{\delta} \right\| \leq 2\|\boldsymbol{\delta}\|^2 \\
\Leftrightarrow \quad & \|\boldsymbol{x}^* + \boldsymbol{\delta} - \|\boldsymbol{x}^* + \boldsymbol{\delta}\|(\boldsymbol{x}^* + \boldsymbol{\delta} - \boldsymbol{x}^*(\boldsymbol{x}^*)^\top\boldsymbol{\delta})\|^2 \leq 4\|\boldsymbol{x}^* + \boldsymbol{\delta}\|^2\|\boldsymbol{\delta}\|^4 \\
\Leftrightarrow \quad & \|(1-u)(\boldsymbol{x}^* + \boldsymbol{\delta}) + u((\boldsymbol{x}^*)^\top\boldsymbol{\delta})\boldsymbol{x}^*\|_2^2 \leq 4u^2v^2 \\
\Leftrightarrow \quad & (1-u)^2u^2 + u^2\Big(\frac{u^2 - v - 1}{2}\Big)^2 + 2u(1-u)\frac{u^2 - v - 1}{2}\frac{u^2 - v + 1}{2} \leq 4u^2v^2 \\
\Leftrightarrow \quad & (3.95).
\end{aligned}
$$

Finally, by the triangle inequality, we have

$$
|\|\boldsymbol{x}^* + \boldsymbol{\delta}\| - \|\boldsymbol{x}^*\|| \leq \|\boldsymbol{\delta}\| \leq \|-\boldsymbol{x}^*\| + \|\boldsymbol{x}^* + \boldsymbol{\delta}\|,
$$

which in turn verifies $(1-u)^2 \leq v \leq (1+u)^2$. This completes our proof of the lemma. □

Next, to extend the result in Lemma 3.5 to any $\boldsymbol{x} \in \mathbb{R} \setminus \{0\}$, we substitute $\boldsymbol{x}^* = \boldsymbol{x}/\|\boldsymbol{x}\|$ and $\boldsymbol{\delta} = \boldsymbol{\delta}/\|\boldsymbol{x}\|$ into (3.93) and obtain

$$
\sup_{\boldsymbol{y} \in \Pi_{\mathcal{C}}(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} + \frac{\boldsymbol{\delta}}{\|\boldsymbol{x}\|})} \left\| \boldsymbol{y} - \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \Big(\boldsymbol{I}_n - \frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|^2}\Big)\frac{\boldsymbol{\delta}}{\|\boldsymbol{x}\|} \right\| \leq 2\frac{\|\boldsymbol{\delta}\|^2}{\|\boldsymbol{x}\|^2}. \tag{3.96}
$$

Since the projection onto the unit sphere is scale-invariant,

$$
\Pi_{\mathcal{C}}\Big(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} + \frac{\boldsymbol{\delta}}{\|\boldsymbol{x}\|}\Big) = \Pi_{\mathcal{C}}(\boldsymbol{x} + \boldsymbol{\delta}). \tag{3.97}
$$

Substituting (3.97) into (3.96) yields (3.6). Thus, by Definition 3.2, for any $\boldsymbol{x} \neq \boldsymbol{0}$ we obtain

$$\nabla \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \frac{1}{\|\boldsymbol{x}\|}\Big(\boldsymbol{I}_n - \frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|^2}\Big), \quad c_1(\boldsymbol{x}) = \infty, \quad c_2(\boldsymbol{x}) = \frac{2}{\|\boldsymbol{x}\|^2}.$$

3.6.8 Details of Application 3.4.2 - Sparse Recovery

3.6.8.1 Proof of (3.42)

In this subsection, we first show that any $\boldsymbol{x}^* \in \Phi_{=s}$ and $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, |x^*_{[s]}|/\sqrt{2})$ share the same index set of s-largest elements (in magnitude), i.e., $\Omega_s(\boldsymbol{x}^*)$. Then, we construct a counter-example to demonstrate that $|x^*_{[s]}|/\sqrt{2}$ is the largest possible radius so that (3.42) holds.

First, we show that for any $i \in \Omega_s(\boldsymbol{x}^*)$ and $j \in \{1, \ldots, n\} \setminus \Omega_s(\boldsymbol{x}^*)$, $|x_j| < |x_i|$ as follows. In particular, we have

$$\begin{aligned} |x_j - x^*_j| + |x_i - x^*_i| &\leq \sqrt{2((x_j - x^*_{[j]})^2 + (x_i - x^*_i)^2)} \\ &\leq \sqrt{2\|\boldsymbol{x} - \boldsymbol{x}^*\|^2} < |x^*_{[s]}|, \end{aligned}$$

where the last inequality stems from the fact that $\|\boldsymbol{x} - \boldsymbol{x}^*\| < |x^*_{[s]}|/\sqrt{2}$. Now,

since $x_j^* = 0$ for all $j \in \{1, \ldots, n\} \setminus \Omega_s(\boldsymbol{x}^*)$, we have

$$\begin{aligned} |x_j| &= |x_j - x_j^*| \\ &< |x_{[s]}^*| - |x_i - x_i^*| \\ &\leq |x_i^*| - |x_i - x_i^*| \\ &\leq |x_i^* + (x_i - x_i^*)| = |x_i|, \end{aligned} \tag{3.98}$$

Therefore, every $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, |x_{[s]}^*|/\sqrt{2})$ shares the same index set of s-largest (in magnitude) elements with $\boldsymbol{x}^*$, i.e., $\Omega_s(\boldsymbol{x}) = \Omega_s(\boldsymbol{x}^*)$, which implies (3.42).

We now construct the counter-example as a point $\boldsymbol{x}$ such that $\Omega_s(\boldsymbol{x}) \neq \Omega_s(\boldsymbol{x}^*)$ and $\boldsymbol{x}$ is not in $\mathcal{B}(\boldsymbol{x}^*, |x_{[s]}^*|/\sqrt{2})$ but arbitrarily close to its boundary. Without loss of generality, assume that $|x_1^*| \geq \ldots \geq |x_s^*| > |x_{s+1}^*| = \ldots = |x_n^*| = 0$. For arbitrarily small $\epsilon > 0$, define $\boldsymbol{x}$ as

$$x_i = \begin{cases} x_s^*/2 & \text{if } i = s, \\ x_s^*/2 + \epsilon & \text{if } i = s+1, \\ x_i & \text{otherwise.} \end{cases}$$

Then, since $x_{s+1} < x_s$, $\boldsymbol{x}$ does not shares the same index set of s-largest (in magnitude) elements with $\boldsymbol{x}^*$. On the other hand, as $\epsilon \to 0$, we have

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| = \sqrt{\sum_{i=1}^{n} (x_i - x_i^*)^2} = \sqrt{\left(-\frac{x_s^*}{2}\right)^2 + \left(\frac{x_s^*}{2} + \epsilon\right)^2} \to \frac{1}{\sqrt{2}} |x_{[s]}^*|.$$

This means $\boldsymbol{x} \notin \mathcal{B}(\boldsymbol{x}^*, |x^*_{[s]}|/\sqrt{2})$ but it can approach the boundary of the ball as ϵ decreases to 0.

3.6.8.2 Proof of Remark 3.6

In the following, we show any stationary point $\boldsymbol{x}^*$ of (3.40) is also a local minimum by proving that the objective function does not decrease if we add any perturbation to $\boldsymbol{x}^*$ on $\mathcal{C}$. Let us consider any perturbation $\boldsymbol{\delta}$ such that $\boldsymbol{\delta} \in \mathcal{B}(\mathbf{0}, c_1(\boldsymbol{x}^*))$ and $\boldsymbol{x} = \boldsymbol{x}^* + \boldsymbol{\delta} \in \mathcal{C}$. Since $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, c_1(\boldsymbol{x}^*))$, using (3.98), we have $|x_{[1]}| \geq \ldots |x_{[s]}| > 0$. On the other hand, since $\boldsymbol{x}$ has no more than s non-zero entries, it must hold that $|x_{[s+1]}| = \ldots = |x_{[n]}| = 0$. Therefore, $\boldsymbol{x} = \boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}_{\boldsymbol{x}^*}^\top \boldsymbol{x}$, which implies $\boldsymbol{\delta} = \boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}_{\boldsymbol{x}^*}^\top \boldsymbol{\delta}$. Now we represent the change in the objective function as

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{A}(\boldsymbol{x}^* + \boldsymbol{\delta}) - \boldsymbol{b}\|^2 - \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\|^2 &= \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{\delta} + \boldsymbol{\delta}^\top \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) \\ &= \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}_{\boldsymbol{x}^*}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}_{\boldsymbol{x}^*}^\top \boldsymbol{\delta} + \boldsymbol{\delta}^\top \boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}_{\boldsymbol{x}^*}^\top \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) \\ &= \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{S}_{\boldsymbol{x}^*}(\boldsymbol{S}_{\boldsymbol{x}^*}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})\boldsymbol{S}_{\boldsymbol{x}^*}^\top \boldsymbol{\delta} \geq 0, \end{aligned} \tag{3.99}$$

where the last equality uses the stationarity condition in (3.43). From (3.99), we conclude $\boldsymbol{x}^*$ is a local minimum of (3.40).

Chapter 4: On Convergence of Projected Gradient Descent for Minimizing a Large-Scale Quadratic over the Unit Sphere[1]

Unit sphere-constrained quadratic optimization has been studied extensively over the past decades. While state-of-art algorithms for solving this problem often rely on relaxation or approximation techniques, there has been little research into scalable first-order methods that tackle the problem in its original form. These first-order methods are often more well-suited for the big data setting. In this chapter, we provide a novel analysis of the simple projected gradient descent method for minimizing a quadratic over a sphere. When the gradient step size is sufficiently small, we show that convergence is locally linear and provide a closed-form expression for the rate. Moreover, a careful selection of the step size can stimulate convergence to the global solution while preventing convergence to local minima.

[1]This work has been published as: Trung Vu, Raviv Raich, and Xiao Fu. "On Convergence of Projected Gradient Descent for Minimizing a Large-Scale Quadratic over the Unit Sphere." In Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6., IEEE, 2019.

4.1 Introduction

This chapter studies the problem of minimizing a quadratic with a norm constraint:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} \qquad \text{subject to } \|\boldsymbol{x}\| \leq 1, \tag{4.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{b} \in \mathbb{R}^n$ and $\|\cdot\|$ is the Euclidean norm.[2] This optimization problem arises frequently in many machine learning and signal processing applications including contour grouping [73], graph partitioning [88] and seismic inversion [149].

If the global solution $\boldsymbol{x}_\star$ of (4.1) lies in the interior of the unit sphere, i.e., $\|\boldsymbol{x}_\star\| < 1$, then $\boldsymbol{x}_\star$ is also the solution of the unconstrained problem. Thus, it is more challenging to consider the case when $\|\boldsymbol{x}_\star\| = 1$. To that end, we restrict our interest to the following problem of minimizing a quadratic over a sphere:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} \qquad \text{subject to } \|\boldsymbol{x}\|^2 = 1, \tag{4.2}$$

In this formulation, we assume that $\boldsymbol{A}$ is symmetric, but not necessarily positive semidefinite. Hence, the objective function is potentially non-convex. Additionally, the norm constraint is non-convex. Both (4.1) and (4.2) are instances of quadratic constrained quadratic program with only one constraint (QCQP-1) and they have been extensively studied in the literature. State-of-art methods for solving this type of QCQP-1 problems in polynomial time include semidefinite relaxation (SDR)

[2]Generally, we can always assume $\boldsymbol{A}$ to be symmetric. Otherwise, one can define an equivalent objective function using $\hat{\boldsymbol{A}} = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^T)$.

[142] and Lagrangian relaxation [165]. However, the problem size for these methods often grows quadratically, making them inapplicable to large-scale problems.

From a different standpoint, problems (4.1) and (4.2) also arise in linear algebra and optimization as the trust-region subproblem. There have been a few extensions to large-scale settings. In [80], Golub and von Matt leveraged the theory of Gauss quadrature and proposed a method to approximately solve (4.1) by tridiagonalizing $\boldsymbol{A}$ using the Lanczos process. In another approach, Sorensen [187] recast the trust-region subproblem in terms of a parameterized eigenvalue problem and developed an implicitly restarted Lanczos method. Related schemes can also be found in [175, 179]. In 2001, Hager introduced sequential subspace method (SSM) [86, 87], carrying out the minimization over a sequence of subspaces that are adjusted after each sequential quadratic programming (SQP) iterate. Similar to the aforementioned methods, SSM relies on Lanczos process to compute the smallest eigenvalue and the corresponding eigenvector of $\boldsymbol{A}$.

In this chapter, we focus on scalable first-order methods for solving (4.2) directly. We leverage the use of simple gradient projection and establish convergence results in the non-convex setting of the spherically constrained quadratic minimization problem, where most convergence guarantees in convex optimization start to break. Our analysis provides a novel insight into behaviors of the algorithm in the neighborhoods of the local optima. Understanding these convergence properties enables us to (i) accommodate acceleration near the optimum—e.g., by using optimal step size selection or momentum methods—and (ii) identify ways of enabling convergence to global solution. Finally, we present numerical results that illustrate

the theory developed in the chapter.

4.2 Solution Properties

Consider the Lagrange function

$$\mathcal{L}(\boldsymbol{x}, \gamma) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} - \frac{1}{2}\gamma(\|\boldsymbol{x}\|^2 - 1)$$

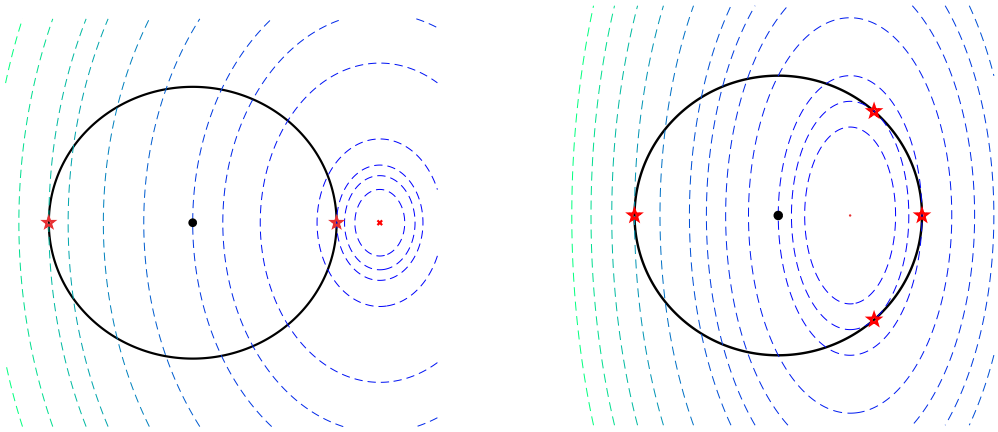
where γ is the Lagrange multiplier. The first-order Lagrangian conditions for optimality can be specified as

$$\begin{cases} \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}, \gamma) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} - \gamma\boldsymbol{x} = \boldsymbol{0}, \\ \nabla_{\gamma}\mathcal{L}(\boldsymbol{x}, \gamma) = \|\boldsymbol{x}\|^2 - 1 = 0. \end{cases}$$

For notational simplicity, we denote the residual by $\boldsymbol{r} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$ and the unit sphere by $\mathcal{S}^{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = 1\}$. Formally, these conditions are given in the following lemma.

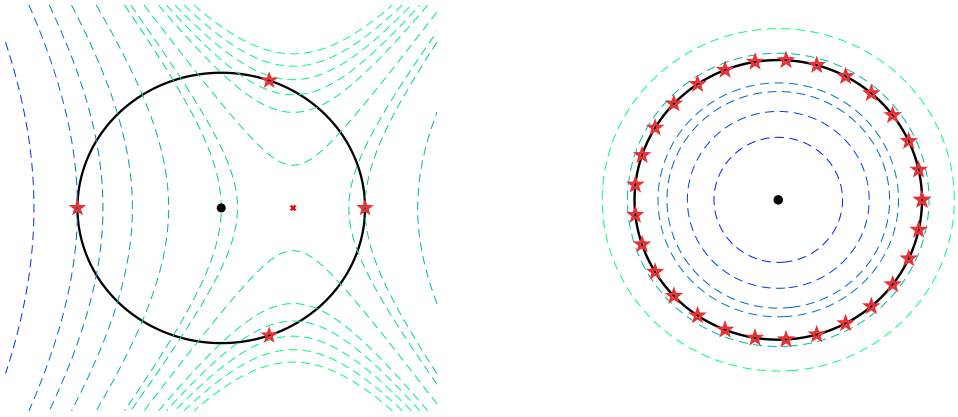
Lemma 4.1 (Stationary conditions)**.** *The vector $\boldsymbol{x}_*$ is a stationary point of problem (4.2) if and only if $\boldsymbol{x}_* \in \mathcal{S}^{n-1}$ and there exists a constant $\gamma(\boldsymbol{x}_*)$ such that $\boldsymbol{r}_* = \boldsymbol{A}\boldsymbol{x}_* - \boldsymbol{b} = \gamma(\boldsymbol{x}_*) \cdot \boldsymbol{x}_*$.*

For the rest of this chapter, we use the shorthand notation γ to refer $\gamma(\boldsymbol{x}_*)$. Lemma 4.1 also implies $\gamma = \boldsymbol{r}_*^T\boldsymbol{x}_*$. Denote $\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp} = \boldsymbol{I} - \boldsymbol{x}_*\boldsymbol{x}_*^T$. Let $\lambda_n = 0$ be the zero eigenvalue corresponding to the eigenvector $\boldsymbol{x}_*$ of the matrix $\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{n-1}$ be the remaining $n-1$ eigenvalues. Noticeably, the



(a) $\boldsymbol{A} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$

(b) $\boldsymbol{A} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$



(c) $\boldsymbol{A} = \begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$

(d) $\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Figure 4.1: Examples of minimizing a quadratic over a sphere. Stationary points are given in red stars. In 2D scenario, they can be either local minima or local maxima (a-c). Furthermore, it is possible that a local optimum lies in a continuum of optima (d).

eigenvalues of $\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}$ can be bounded by

$$\lambda_{\min}(\boldsymbol{A}) \leq \lambda_{n-1} \leq \ldots \leq \lambda_1 \leq \lambda_{\max}(\boldsymbol{A}),$$

where $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ are the smallest and largest eigenvalues of $\boldsymbol{A}$, respectively. Moreover, the relationship among those eigenvalues and the Lagrange multiplier provides necessary and sufficient conditions for determining the type of a stationary point.

Lemma 4.2. *A stationary point $\boldsymbol{x}_*$ of problem (4.2) is a strict local minimum if and only if $\gamma(\boldsymbol{x}_*) < \lambda_{n-1}(\boldsymbol{x}_*)$. Furthermore, $\boldsymbol{x}_\star$ is a **global minimizer** of problem (4.2) if and only if $\gamma(\boldsymbol{x}_\star) \leq \lambda_{\min}(\boldsymbol{A})$.*

Example 4.1. *Figure 4.1 demonstrates various cases where there are different numbers of stationary points. As an exemplification, let us examine the derivation of the problem in Fig. 4.1(b):*

$$\min_{x_1,x_2} \frac{1}{2}(4x_1^2 + x_2^2) - 2x_1 \qquad s.t.\ x_1^2 + x_2^2 = 1.$$

For each stationary point $\boldsymbol{x}_*$, the matrix $\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}$ has one zero eigenvalue corresponding to the eigenvector $\boldsymbol{x}_*$, and the other non-zero eigenvalue $\lambda_1 = \lambda_{n-1}$ (since $n = 2$) lies between $\lambda_{\min}(\boldsymbol{A}) = 1$ and $\lambda_{\max}(\boldsymbol{A}) = 4$. Omitting the detailed calculation, we list the four stationary points of this problem as follows: (i) a global maximum at $[x_1, x_2] = [-1, 0]$ with $\gamma = 6, \lambda_1 = 1$; (ii) a local maximum at $[x_1, x_2] = [1, 0]$ with $\gamma = 2, \lambda_1 = 1$; and (iii) 2 local (also global) minima at

$[x_1, x_2] = [\frac{2}{3}, \pm\frac{\sqrt{5}}{3}]$ *with* $\gamma = 1, \lambda_1 = \frac{8}{3}$.

4.3 The Projected Gradient Algorithm

The projected gradient descent approach (see Algorithm 4.1) starts at an initial point $x^{(0)}$, then performs the update

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{f}_\alpha(\boldsymbol{x}^{(t)}) = \mathcal{P}_{\mathcal{S}^{n-1}}\Big(\boldsymbol{x}^{(t)} - \alpha(\boldsymbol{A}\boldsymbol{x}^{(t)} - \boldsymbol{b})\Big), \tag{4.3}$$

where $\alpha > 0$ is the step size and $\mathcal{P}_{\mathcal{S}^{n-1}}(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is the spherical projection uniquely given by

$$\mathcal{P}_{\mathcal{S}^{n-1}}(\boldsymbol{x}) = \begin{cases} \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} & \text{if } \boldsymbol{x} \neq \boldsymbol{0}, \\ \boldsymbol{e} & \text{if } \boldsymbol{x} = \boldsymbol{0}, \end{cases}$$

with $\boldsymbol{e} \in \mathcal{S}^{n-1}$ such that $\boldsymbol{e}$ and $\boldsymbol{A}\boldsymbol{e} - \boldsymbol{b}$ are not collinear. The definition of projection at $\boldsymbol{0}$ is just for numerical issues when the algorithm encounters the origin at some iteration. In practice, we can choose $\boldsymbol{e}$ to be one of the natural basis, i.e., $[1, 0, \ldots, 0]$. Next, let us consider some important properties associated with Algorithm 4.1.

Definition 4.1. *A fixed point of* $\boldsymbol{f}_\alpha$ *is defined as any vector* $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ *such that*

$$\boldsymbol{f}_\alpha(\bar{\boldsymbol{x}}) = \mathcal{P}_{\mathcal{S}^{n-1}}\big(\bar{\boldsymbol{x}} - \alpha(\boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b})\big) = \bar{\boldsymbol{x}}.$$

Lemma 4.3. *The vector $\bar{\boldsymbol{x}}$ is a fixed point of $\boldsymbol{f}_\alpha$ if and only if $\bar{\boldsymbol{x}} \in \mathcal{S}^{n-1}$ and there exists a constant $\gamma < \frac{1}{\alpha}$ such that $\bar{\boldsymbol{r}} = \boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b} = \gamma\bar{\boldsymbol{x}}$.*

Proof. Since $\bar{\boldsymbol{x}}$ is a fixed point of $\boldsymbol{f}_\alpha$, we have

$$\bar{\boldsymbol{x}} = \boldsymbol{f}_\alpha(\bar{\boldsymbol{x}}) = \mathcal{P}_{\mathcal{S}^{n-1}}\big(\bar{\boldsymbol{x}} - \alpha(\boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b})\big).$$

Consequently, $\bar{\boldsymbol{x}} \in \mathcal{S}^{n-1}$. Furthermore, if $\bar{\boldsymbol{x}} - \alpha(\boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b}) = \boldsymbol{0}$, then $\bar{\boldsymbol{x}} = \mathcal{P}_{\mathcal{S}^{n-1}}(\boldsymbol{0}) = \boldsymbol{e}$. But this contradicts with the non-collinearity of $\boldsymbol{e}$ and $\boldsymbol{A}\boldsymbol{e} - \boldsymbol{b}$. Thus, it must be the case that $\bar{\boldsymbol{x}} - \alpha(\boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b}) \neq \boldsymbol{0}$, and hence $\bar{\boldsymbol{x}} = \frac{\bar{\boldsymbol{x}} - \alpha\bar{\boldsymbol{r}}}{\|\bar{\boldsymbol{x}} - \alpha\bar{\boldsymbol{r}}\|}$. There exists a constant γ such that $\bar{\boldsymbol{r}} = \gamma\bar{\boldsymbol{x}}$. Substituting back into the fixed-point equation yields

$$\bar{\boldsymbol{x}} = \frac{(1-\alpha\gamma)\bar{\boldsymbol{x}}}{|1-\alpha\gamma|\,\|\bar{\boldsymbol{x}}\|} = \frac{1-\alpha\gamma}{|1-\alpha\gamma|}\frac{\bar{\boldsymbol{x}}}{\|\bar{\boldsymbol{x}}\|} = \operatorname{sign}(1-\alpha\gamma)\cdot\bar{\boldsymbol{x}}.$$

Therefore, the sign of $1-\alpha\gamma$ must be 1 or $1-\alpha\gamma > 0$. □

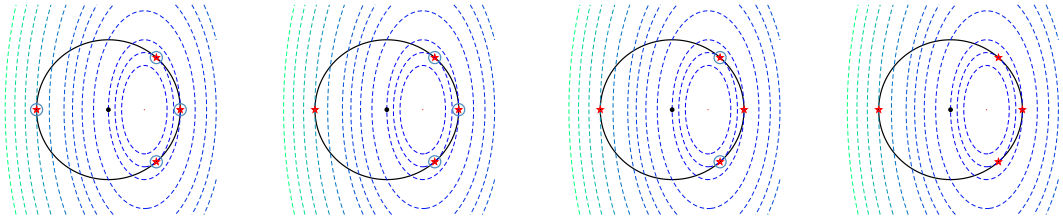
From Lemma 4.2 and 4.3, we can establish the necessary and sufficient conditions for a fixed point of $\boldsymbol{f}_\alpha$ to be a stationary point as follows.

Corollary 4.1. *The vector $\boldsymbol{x}_*$ is a stationary point of problem (4.2) if and only if there exists $\alpha > 0$ such that $\boldsymbol{x}_*$ is a fixed point of $\boldsymbol{f}_\alpha$.*

Example 4.2. *Continued from Example 4.1, we illustrate fixed points with different step sizes in Fig. 4.2. When α is small enough, all stationary points can be fixed points. As α increases, only stationary points with the multiplier $\gamma < 1/\alpha$ remains to be fixed points of $\boldsymbol{f}_\alpha$. Interestingly, while any convergence point of Algorithm 4.1 with step size α is a fixed point of the iterated function $\boldsymbol{f}_\alpha$, the vice*

Algorithm 4.1 Projected Gradient Descent (PGD)

1: Initialize $\boldsymbol{x}^{(0)} \in \mathcal{S}^{n-1}$
2: **for** $t = 0, 1, \ldots$ **do**
3: $\quad \boldsymbol{z}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha(\boldsymbol{A}\boldsymbol{x}^{(t)} - \boldsymbol{b})$
4: $\quad \boldsymbol{x}^{(t+1)} = \mathcal{P}_{\mathcal{S}^{n-1}}(\boldsymbol{z}^{(t+1)})$



(a) $0 < \alpha < 1/6$ (b) $1/6 < \alpha < 1/2$ (c) $1/2 < \alpha < 1$ (d) $\alpha > 1$

Figure 4.2: Stationary points (red stars) versus fixed points (blue circles) with different step size α in optimizing the quadratic objective $\frac{1}{2}(4x_1^2 + x_2^2) - 2x_1$ over the unit circle. Dashed lines are the contour levels of the objective value.

versa is not true: the global maximum at $[x_1, x_2] = [-1, 0]$ *is a fixed point of* $\boldsymbol{f}_\alpha$*, for* $\alpha < 1/6$*, but as can be seen in the next section, it is not a convergence point of the algorithm.*

4.4 Convergence Analysis

In this section, we present our result on the local uniform convergence of Algorithm 4.1 with a certain choice of step size to a strict local optimum. The convergence is shown to be linear and the asymptotic rate is given in a closed-form expression. The challenges come the non-convexity of the norm constraint and (potentially) the negative curvature of the objective function. Let us begin with the analysis of the projection operator.

Lemma 4.4 (Taylor series expansion of the projection)**.** *Let $\boldsymbol{x} \in \mathbb{R}^n$ be a nonzero vector and $\boldsymbol{\delta}$ be a small perturbation such that $\|\boldsymbol{\delta}\| \ll \|\boldsymbol{x}\|$. Then,*

$$\mathcal{P}_{\mathcal{S}^{n-1}}(\boldsymbol{x}+\boldsymbol{\delta}) = \mathcal{P}_{\mathcal{S}^{n-1}}(\boldsymbol{x}) + \frac{1}{\|\boldsymbol{x}\|}\Big(\boldsymbol{I} - \frac{\boldsymbol{x}\boldsymbol{x}^T}{\|\boldsymbol{x}\|^2}\Big)\boldsymbol{\delta} + O\big(\|\boldsymbol{\delta}\|^2\big).$$

Now, considering the convergence of Algorithm 4.1 in the region near a strict local minimum $\boldsymbol{x}_*$ where $\boldsymbol{r}_* = \boldsymbol{A}\boldsymbol{x}_* - \boldsymbol{b} = \gamma \boldsymbol{x}_*$. Denote $\boldsymbol{\delta}^{(t)} = \boldsymbol{x}^{(t)} - \boldsymbol{x}_*$ and reorganize the update equation (4.3) as

$$\begin{aligned}\boldsymbol{\delta}^{(t+1)} &= \mathcal{P}_{\mathcal{S}^{n-1}}\big(\boldsymbol{x}_* - \alpha(\boldsymbol{A}\boldsymbol{x}_* - \boldsymbol{b}) + (\boldsymbol{I} - \alpha\boldsymbol{A})\boldsymbol{\delta}^{(t)}\big) - \boldsymbol{x}_* \\ &= \mathcal{P}_{\mathcal{S}^{n-1}}\big((1-\alpha\gamma)\boldsymbol{x}_* + (\boldsymbol{I} - \alpha\boldsymbol{A})\boldsymbol{\delta}^{(t)}\big) - \mathcal{P}_{\mathcal{S}^{n-1}}\big((1-\alpha\gamma)\boldsymbol{x}_*\big).\end{aligned}$$

Substituting $\boldsymbol{x} = (1-\alpha\gamma)\boldsymbol{x}_*$ and $\boldsymbol{\delta} = (\boldsymbol{I} - \alpha\boldsymbol{A})\boldsymbol{\delta}^{(t)}$ into Lemma 4.4 yields

$$\boldsymbol{\delta}^{(t+1)} = \frac{1}{\|(1-\alpha\gamma)\boldsymbol{x}_*\|}\big(\boldsymbol{I} - \frac{(1-\alpha\gamma)\boldsymbol{x}_*(1-\alpha\gamma)\boldsymbol{x}_*^T}{\|(1-\alpha\gamma)\boldsymbol{x}_*\|^2}\big)(\boldsymbol{I} - \alpha\boldsymbol{A})\boldsymbol{\delta}^{(t)} + O\big(\|\boldsymbol{\delta}^{(t)}\|^2\big).$$

Assume the step size is chosen such that $\alpha\gamma < 1$, and recall that $\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp} = \boldsymbol{I} - \boldsymbol{x}_*\boldsymbol{x}_*^T$ for $\|\boldsymbol{x}_*\| = 1$. The recursion can be rewritten as

$$\boldsymbol{\delta}^{(t+1)} = \frac{1}{1-\alpha\gamma}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}(\boldsymbol{I} - \alpha\boldsymbol{A})\boldsymbol{\delta}^{(t)} + O\big(\|\boldsymbol{\delta}^{(t)}\|^2\big). \tag{4.4}$$

The stability of a general nonlinear difference equation of the form $x^{k+1} = Tx^k + o(\|x^k\|)$ has been well-studied in [15, 166]. In particular, let ρ_α be the spectral radius of $(1-\alpha\gamma)^{-1}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}(\boldsymbol{I} - \alpha\boldsymbol{A})$, i.e., the largest absolute value of its eigenvalues.

If $\rho_\alpha < 1$, then the series $\{\boldsymbol{\delta}^{(t)}\}$ approaches zeros with sufficiently small $\boldsymbol{\delta}^{(0)}$, where

$$\|\boldsymbol{\delta}^{(t)}\| \leq K\|\boldsymbol{\delta}^{(0)}\|\big(\rho_\alpha + o(1)\big)^t.$$

Therefore, to prove the local convergence of the PGD algorithm, it is sufficient to show that $\rho_\alpha < 1$. We present our main result on the local uniform convergence to a strict local minimum as follows.

Definition 4.2. *Algorithm 4.1 with a fixed step size α is said to converges* ***locally uniformly*** *to $\boldsymbol{x}_*$ if and only if there exists a constant ϵ such that for any $\boldsymbol{x}^{(0)}$ satisfying $\big\|\boldsymbol{x}^{(0)} - \boldsymbol{x}_*\big\| \leq \epsilon$, we have $\big\|\boldsymbol{x}^{(t)} - \boldsymbol{x}_*\big\| \leq \epsilon$, for all $t = 0, 1, \ldots$ and $\lim_{t\to\infty} \big\|\boldsymbol{x}^{(t)} - \boldsymbol{x}_*\big\| = 0$.*

Theorem 4.1. *The vector $\boldsymbol{x}_*$ is a strict local minimum of problem (4.2) such that $\gamma < \lambda_{n-1}$ if and only if there exists $\alpha > 0$ such that Algorithm 4.1 with step size α converges locally uniformly to $\boldsymbol{x}_*$. Furthermore, for any step size $\alpha > 0$ such that $\alpha(\lambda_1 + \gamma) < 2$, the sequence $\{\boldsymbol{x}^{(t)}\}$ satisfies*

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{x}_*\| \leq K\|\boldsymbol{x}^{(0)} - \boldsymbol{x}_*\|\big(\rho_\alpha + o(1)\big)^t,$$

for some constant $K > 0$ and $\rho_\alpha = \max_{1\leq i\leq n-1} \frac{|1-\alpha\lambda_i|}{1-\alpha\gamma}$.

The proof of Theorem 4.1 is given in the appendix. The theorem reveals PGD converges to a local minimum at an asymptotic linear rate ρ_α. Note that in our problem, $\boldsymbol{A}$ is not necessarily PSD, meaning λ_i could be negative. To facilitate acceleration, one can speed up the convergence by optimizing over the step size α.

Lemma 4.5. *The optimal rate of local convergence and the optimal step size for Algorithm 4.1 are given by* [3]

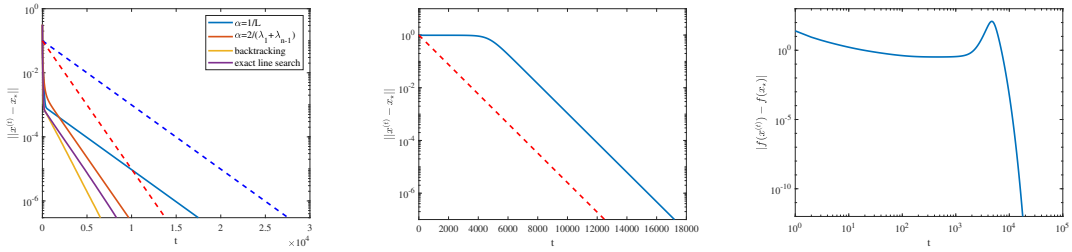
$$\begin{cases} \rho_* = \frac{\lambda_1 - \lambda_{n-1}}{\lambda_1 + \lambda_{n-1} - 2\gamma}, \quad \alpha_* = \frac{2}{\lambda_1 + \lambda_{n-1}} & \text{if } \lambda_1 + \lambda_{n-1} > 0 \\ \rho_* = \frac{\lambda_{n-1}}{\gamma}, \qquad\qquad \alpha_* = \infty & \text{otherwise.} \end{cases}$$

Example 4.3. *Continued from Example 4.2, Theorem 4.1 states that the PGD algorithm only converges locally uniformly to the two local minima at* $[2/3, \pm\sqrt(5)/3]$ *with* $\lambda_1 = \frac{8}{3}$. *Notice that these points are fixed points of* $\boldsymbol{f}_\alpha$ *for* $\alpha < 1$. *Since* $\lambda_1 = \lambda_{n-1}$, *the optimal rate is* $\rho_* = 0$ *with step size* $\alpha_* = \frac{3}{8}$. *In this case, the convergence is quadratic due to the residual term in (4.4).*

Global convergence of the PGD algorithm. Theorem 4.1 also implies that for any step size α satisfying $\rho_\alpha > 1$, the algorithm tends to move away from the local minimum $\boldsymbol{x}_*$. This intuition leads us to the following strategy for step size selection: choosing α large enough such that $g(\alpha, \boldsymbol{x}_*) \geq 2$, where $g(\alpha, \boldsymbol{x}_*) \triangleq \alpha(\lambda_1(\boldsymbol{x}_*) + \gamma(\boldsymbol{x}_*))$, for all strict local minimum $\boldsymbol{x}_*$ except the global minimum $\boldsymbol{x}_\star$.

Remark 4.1. *Assume that there exists sufficiently large* α *satisfying* $g(\alpha, \boldsymbol{x}_\star) < 2$ *for any global minimum* $\boldsymbol{x}_\star$ *and* $g(\alpha, \boldsymbol{x}_*) \geq 2$ *for any strict local minimum* $\boldsymbol{x}_*$. *Then Algorithm 4.1 with step size* α *converges to one of the optimal solutions* $\boldsymbol{x}_\star$ *at an asymptotic geometric rate of* $\rho_\alpha(\boldsymbol{x}_\star)$.

[3] The proofs of Lemma 4.2 and Lemma 4.5 are given in the Appendix at the end of this chapter.



(a) (b) (c)

Figure 4.3: (a) Local convergence of the projected gradient method with different step sizes for solving a unit-constrained least squares. (b-c) Empirical evidence of the convergence to global optimum where $\boldsymbol{x}^{(0)}$ is initialized near a local optimum and α is chosen according to Remark 4.1. Dashed lines are added as an illustration for the theoretical bounds for the convergence rate of fixed step size methods (up to a constant).

4.5 Numerical Results

Motivating applications of problem (4.1) are the well-known trust-region subproblem in nonlinear optimization [49], the variational problem in structural limit analysis [74], and the optimizing precoder method in transmitter based CDMA optimization [97]. For the purpose of demonstrating the theoretical analysis, we will focus on numerical results for local convergence of Algorithm 4.1 with different step sizes, and empirical evidence for our conjecture about the global convergence with an appropriate step size. In our experiment, we first generate a random symmetric matrix $\boldsymbol{A}$ of size $n = 1000$ such that the smallest eigenvalue is far away from the other eigenvalues. Then, we choose one multiplier for the global solution $\gamma(\boldsymbol{x}_\star) < \lambda_{\min}(\boldsymbol{A})$ and one multiplier for the local solution $\gamma(\boldsymbol{x}_*) > \lambda_{\min}(\boldsymbol{A})$. Next, the coefficient vector $\boldsymbol{b}$ is chosen such that $\boldsymbol{b}^T(\boldsymbol{A} - \gamma(\boldsymbol{x}_\star)\boldsymbol{I})^{-2}\boldsymbol{b} = \boldsymbol{b}^T(\boldsymbol{A} - \gamma(\boldsymbol{x}_*)\boldsymbol{I})^{-2}\boldsymbol{b} = 1$. Finally, we compute $\boldsymbol{x}_\star = (\boldsymbol{A} - \gamma(\boldsymbol{x}_\star)\boldsymbol{I})^{-1}\boldsymbol{b}$ and $\boldsymbol{x}_* = (\boldsymbol{A} - \gamma(\boldsymbol{x}_*)\boldsymbol{I})^{-1}\boldsymbol{b}$.

For local convergence, starting at an initial point $\boldsymbol{x}^{(0)}$ close to the local minimum $\boldsymbol{x}_*$, we examine four PGD algorithms with different step sizes:

(i) Commonly used step size: $\alpha = 1/L$, where L is the spectral radius of $\boldsymbol{A}$. This step size selection is often used in many classic proofs of convergence in convex optimization, where an L-smooth objective function can be guaranteed to monotonically decrease through PGD iterations.

(ii) Optimal step size: $\alpha = \alpha_*$ as in Lemma 4.5. We choose $\boldsymbol{A}$ and $\boldsymbol{x}_*$ such that $\lambda_1 + \lambda_{n-1} > 0$, hence $\alpha_* = 2/(\lambda_1 + \lambda_{n-1})$.

(iii) Projected backtracking line search: rewriting the PGD update as generalized gradient step $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t \boldsymbol{G}_{\alpha_t}(\boldsymbol{x}^{(t)})$, where $\boldsymbol{G}_\alpha(\boldsymbol{x}) = \frac{1}{\alpha}\Big(\boldsymbol{x} - \mathcal{P}_{\mathcal{S}^{n-1}}\big(\boldsymbol{x} - \alpha(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\big)\Big)$. Denote the quadratic objective by $q(\boldsymbol{x})$. Starting with $\alpha = 1$, we shrink $\alpha = \beta\alpha$, for $0 < \beta < 1$, while

$$q\big(\boldsymbol{x} - \alpha \boldsymbol{G}_\alpha(\boldsymbol{x})\big) > q(\boldsymbol{x}) - \alpha(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T \boldsymbol{G}_\alpha(\boldsymbol{x}) + \frac{\alpha}{2} \left\| \boldsymbol{G}_\alpha(\boldsymbol{x}) \right\|^2 .$$

In our case, this backtracking condition can be simplified to

$$\boldsymbol{G}_\alpha(\boldsymbol{x})^T \boldsymbol{A} \boldsymbol{G}_\alpha(\boldsymbol{x}) > \frac{1}{\alpha} \left\| \boldsymbol{G}_\alpha(\boldsymbol{x}) \right\|^2 .$$

(iv) Exact line search: finding the step size that maximizes the decrease in objective function

$$\alpha_{\min} = \min_{\alpha > 0} q\big(\mathcal{P}_{\mathcal{S}^{n-1}}(\boldsymbol{x} - \alpha(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}))\big).$$

As can be seen from Fig. 4.3(a), the convergence of PGD with step size $\alpha = 1/L$ (blue) is the slowest among the considered methods, followed by the one

with optimal step size $\alpha = 2/(\lambda_1 + \lambda_{n-1})$ (red). Note that they both match the asymptotic rate predicted in theory. The adaptive schemes, namely projected backtracking (yellow) and exact line search (magenta) perform slightly better than the optimal fixed step size scheme.

For global convergence, we purposely initialize the algorithm at the same $\boldsymbol{x}^{(0)}$ that is close to the **local** minimum $\boldsymbol{x}_*$, and run the PGD algorithm with step size $\alpha = \frac{1}{2}\big(\frac{2}{\lambda_1(\boldsymbol{x}_\star)+\gamma(\boldsymbol{x}_\star)} + \frac{2}{\lambda_1(\boldsymbol{x}_*)+\gamma(\boldsymbol{x}_*)}\big)$. It is easy to verify that α satisfies the condition in Remark 4.1: $g(\alpha, \boldsymbol{x}_*) < 2 < g(\alpha, \boldsymbol{x}_\star)$. Figure 4.3 demonstrates the convergence of the algorithm to the global minimizer $\boldsymbol{x}_\star$, in terms of the distance to the solution (b) and the decrease in the objective value (c). In the first 5000 iterations, the algorithm tries to escape from the local minimum. Then it experiences a period of fluctuation before getting attracted by the global minimum. Notice that when it reaches the neighborhood of $\boldsymbol{x}_\star$, (monotonic) linear convergence is observed. [4]

4.6 Conclusion and Future Work

We analyzed the projected gradient descent approach to minimizing a quadratic over a sphere. We showed that the algorithm always converges linearly to a strict local minimum in its neighborhood. Further, we provided the closed-form expression for convergence rate and identified ways of achieving optimal rate of convergence near the optimum. Our analysis can be extended in the following directions: (i)

[4] This result is provided merely as an illustration of a typical run, not to be considered as an empirical proof. In our experiments, we re-ran simulations multiple times with various problem sizes and always observed convergence.

minimizing a quadratic over an ellipsoid; (ii) acceleration of gradient projection using momentum; and (iii) analysis of convergence to a continuum of optima.

4.7 Appendix

4.7.1 Proof of Lemma 2

Recall our optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} \qquad \text{subject to } \|\boldsymbol{x}\|^2 = 1,$$

The proof of the global minimizers is given by Lemmas 2.4 and 2.8 in [186]. Below we provide the proof of the sufficient condition for strict local minima of problem (4.2). This is a consequence of the second-order sufficient condition for optimality in constrained optimization (see Chapter 3 - [16]). Notice that in our case, the Hessian of the Lagrange function is $\nabla^2_{\boldsymbol{x}\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\gamma) = \boldsymbol{A} - \gamma\boldsymbol{I}$ and the Jacobian of the constraint $\boldsymbol{x}^T\boldsymbol{x} - 1 = 0$ is $\boldsymbol{J}(\boldsymbol{x}) = \boldsymbol{x}$. Let $\boldsymbol{x}_*$ be a stationary point of problem (4.2). Then $\boldsymbol{x}_*$ is a strict local minimum if

$$\boldsymbol{y}^T(\boldsymbol{A} - \gamma\boldsymbol{I})\boldsymbol{y} > 0 \quad \forall\ \boldsymbol{y} \text{ s.t. } \boldsymbol{y} \perp \boldsymbol{x}_* \text{ (i.e. } \boldsymbol{y}^T\boldsymbol{x}_* = 0). \tag{4.5}$$

Since $\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{y} = \boldsymbol{y}$ for all $\boldsymbol{y} \perp \boldsymbol{x}$, we have

$$\begin{aligned}\boldsymbol{y}^T(\boldsymbol{A} - \gamma\boldsymbol{I})\boldsymbol{y} &= \boldsymbol{y}^T\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}(\boldsymbol{A} - \gamma\boldsymbol{I})\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{y} \\ &= \boldsymbol{y}^T\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{y} - \gamma\boldsymbol{y}^T\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{y} \\ &= \boldsymbol{y}^T(\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp} - \gamma\boldsymbol{I})\boldsymbol{y}.\end{aligned}$$

Thus, condition (4.5) is equivalent to $\boldsymbol{y}^T(\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp} - \gamma\boldsymbol{I})\boldsymbol{y} > 0$, or

$$\gamma < \frac{\boldsymbol{y}^T\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{y}}{\|\boldsymbol{y}\|^2} \quad \forall\ \boldsymbol{y} \text{ s.t. } \boldsymbol{y} \perp \boldsymbol{x}_*. \tag{4.6}$$

On the other hand, by the definition of λ_{n-1}, we have

$$\lambda_{n-1} = \min_{\boldsymbol{y}\perp\boldsymbol{x}_*} \frac{\boldsymbol{y}^T\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{y}}{\|\boldsymbol{y}\|^2} = \min_{\boldsymbol{y}\perp\boldsymbol{x}_*} \frac{\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}}{\|\boldsymbol{y}\|^2} = \min_{\substack{\boldsymbol{y}\perp\boldsymbol{x}_* \\ \|\boldsymbol{y}\|=1}} \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}. \tag{4.7}$$

Combining (4.5), (4.6) and (4.7), we conclude $\gamma < \lambda_{n-1}$ implies $\boldsymbol{x}_*$ is a strict local minimum of problem (4.2).

It is noteworthy that the necessary condition for local minima of problem (4.2), following a similar argument, is given by $\gamma \leq \lambda_{n-1}$. However, it is possible that a strict local minimum associates with $\gamma = \lambda_{n-1}$. For example, consider the 2D-case

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \boldsymbol{x}_* = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix}, \qquad \gamma = \lambda_{n-1} = 1.$$

It can be seen that the curvature of the objective function almost coincides with

that of the unit sphere at $\boldsymbol{x}_*$ in the above example. The following lemma states the necessary condition for strict local minima of problem (4.2):

Lemma 4.6. *If $\boldsymbol{x}_*$ is a strict local minimum of problem (4.2), then either of the following condition holds*

- $\gamma < \lambda_{n-1}$
- $\boldsymbol{x}_*^T \boldsymbol{A} \boldsymbol{x}_* > \boldsymbol{u}^T \boldsymbol{A} \boldsymbol{u} = \gamma = \lambda_{n-1}$ *and* $\boldsymbol{x}_*^T \boldsymbol{A} \boldsymbol{u} = 0$ *for* $\boldsymbol{u} = \underset{\substack{\boldsymbol{y} \perp \boldsymbol{x}_* \\ \|\boldsymbol{y}\|=1}}{\operatorname{argmin}} \boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}$.

Proof. By definition of strict local minima, for any $\boldsymbol{x} \in \mathcal{S}^{n-1}$ such that $0 < \|\boldsymbol{x} - \boldsymbol{x}_*\| < \epsilon$ with sufficiently small $\epsilon > 0$, we have

$$
\begin{aligned}
0 < & f(\boldsymbol{x}) - f(\boldsymbol{x}_*) \\
&= \Big(\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} - \boldsymbol{b}^T \boldsymbol{x}\Big) - \Big(\frac{1}{2}\boldsymbol{x}_*^T \boldsymbol{A} \boldsymbol{x}_* - \boldsymbol{b}^T \boldsymbol{x}_*\Big) \\
&= \Big(\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} - (\boldsymbol{A}\boldsymbol{x}_* - \gamma \boldsymbol{x}_*)^T \boldsymbol{x}\Big) - \Big(\frac{1}{2}\boldsymbol{x}_*^T \boldsymbol{A} \boldsymbol{x}_* - (\boldsymbol{A}\boldsymbol{x}_* - \gamma \boldsymbol{x}_*)^T \boldsymbol{x}_*\Big) \\
&\qquad (\text{since } \boldsymbol{A}\boldsymbol{x}_* - \boldsymbol{b} = \gamma \boldsymbol{x}_*) \\
&= \frac{1}{2}\Big(\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} - 2\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}_* + \boldsymbol{x}_*^T \boldsymbol{A} \boldsymbol{x}_* - \gamma(2\boldsymbol{x}_*^T \boldsymbol{x}_* - 2\boldsymbol{x}^T \boldsymbol{x}_*)\Big) \\
&= \frac{1}{2}\Big((\boldsymbol{x} - \boldsymbol{x}_*)^T \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{x}_*) - \gamma \|\boldsymbol{x} - \boldsymbol{x}_*\|^2\Big) \qquad (\text{since } \|\boldsymbol{x}\| = \|\boldsymbol{x}_*\| = 1)
\end{aligned}
$$

Denote $\boldsymbol{\delta} = \boldsymbol{x} - \boldsymbol{x}_* = \boldsymbol{\delta}_x + \boldsymbol{\delta}_\perp$, where $\boldsymbol{\delta}_x$ is collinear to $\boldsymbol{x}_*$ and $\boldsymbol{\delta}_\perp$ is orthogonal to $\boldsymbol{x}_*$. The last inequality becomes

$$
\gamma < \frac{\boldsymbol{\delta}^T \boldsymbol{A} \boldsymbol{\delta}}{\|\boldsymbol{\delta}\|^2} = \frac{\boldsymbol{\delta}_x^T \boldsymbol{A} \boldsymbol{\delta}_x + 2\boldsymbol{\delta}_x^T \boldsymbol{A} \boldsymbol{\delta}_\perp + \boldsymbol{\delta}_\perp \boldsymbol{A} \boldsymbol{\delta}_\perp}{\|\boldsymbol{\delta}\|^2}. \tag{4.8}
$$

Using the fact that $\|\boldsymbol{\delta}\|^2 = \|\boldsymbol{\delta}_x\|^2 + \|\boldsymbol{\delta}_\perp\|^2$ and

$$1 = \|\boldsymbol{x}\|^2 = \|\boldsymbol{x}_* + \boldsymbol{\delta}\|^2 = \|\boldsymbol{x}_*\|^2 + \|\boldsymbol{\delta}\|^2 + 2\boldsymbol{x}_*^T\boldsymbol{\delta} = 1 + \|\boldsymbol{\delta}_x\|^2 + \|\boldsymbol{\delta}_\perp\|^2 + 2\boldsymbol{x}_*^T\boldsymbol{\delta}_x$$

we obtain $\boldsymbol{\delta}_x = -\|\boldsymbol{\delta}_x\|\,\boldsymbol{x}_*$ and $\|\boldsymbol{\delta}_\perp\| = \sqrt{2\|\boldsymbol{\delta}_x\| - \|\boldsymbol{\delta}_x\|^2}$. Substituting back into (4.8) yields

$$\begin{aligned}\gamma &< \frac{\|\boldsymbol{\delta}_x\|^2\,\boldsymbol{x}_*^T\boldsymbol{A}\boldsymbol{x}_* - 2\|\boldsymbol{\delta}_x\|\sqrt{2\|\boldsymbol{\delta}_x\| - \|\boldsymbol{\delta}_x\|^2}\,\boldsymbol{x}_*^T\boldsymbol{A}\boldsymbol{u} + (2\|\boldsymbol{\delta}_x\| - \|\boldsymbol{\delta}_x\|^2)\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u}}{2\|\boldsymbol{\delta}_x\|} \\ &= \frac{1}{2}\Big(\|\boldsymbol{\delta}_x\|\,\boldsymbol{x}_*^T\boldsymbol{A}\boldsymbol{x}_* - 2\sqrt{2\|\boldsymbol{\delta}_x\| - \|\boldsymbol{\delta}_x\|^2}\,\boldsymbol{x}_*^T\boldsymbol{A}\boldsymbol{u} - \|\boldsymbol{\delta}_x\|\,\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u}\Big) + \boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u}, \quad (4.9)\end{aligned}$$

where $\boldsymbol{u}$ is the unit-length vector that is collinear to $\boldsymbol{\delta}_\perp$. Now since $\|\boldsymbol{\delta}_x\|$ can be chosen arbitrarily small and $\boldsymbol{u}$ can be chosen in any direction that is orthogonal to $\boldsymbol{x}_*$, taking $\|\boldsymbol{\delta}_x\| \to 0$ in (4.9) yields $\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} \geq \gamma$ for any unit-length vector $\boldsymbol{u} \perp \boldsymbol{x}_*$. Thus, from (4.7), we conclude that $\lambda_{n-1} \geq \gamma$. Furthermore, if $\lambda_{n-1} = \boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} = \gamma$, then it holds that

$$\|\boldsymbol{\delta}_x\|\,\boldsymbol{x}_*^T\boldsymbol{A}\boldsymbol{x}_* - 2\sqrt{2\|\boldsymbol{\delta}_x\| - \|\boldsymbol{\delta}_x\|^2}\,\boldsymbol{x}_*^T\boldsymbol{A}\boldsymbol{u} - \|\boldsymbol{\delta}_x\|\,\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} > 0 \quad \text{for all } \|\boldsymbol{\delta}_x\|. \tag{4.10}$$

Notice that if $\boldsymbol{x}_*^T\boldsymbol{A}\boldsymbol{u} > 0$, we can always choose sufficiently small $\|\boldsymbol{\delta}_x\|$ so that the second term ($\mathcal{O}(\|\boldsymbol{\delta}_x\|^{1/2})$) on the LHS of (4.10) dominates the other terms ($\mathcal{O}(\|\boldsymbol{\delta}_x\|)$), which in turn forces the LHS to be negative. Otherwise, if $\boldsymbol{x}_*^T\boldsymbol{A}\boldsymbol{u} < 0$, we can replace $\boldsymbol{u}$ by $-\boldsymbol{u}$ and follows the same argument to expose the contradiction. Therefore, it must hold that $\boldsymbol{x}_*^T\boldsymbol{A}\boldsymbol{u} = 0$ in the case $\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} = \gamma$. In addition,

substituting these quantities back into (4.10) yields $\boldsymbol{x}_*^T \boldsymbol{A}\boldsymbol{x}_* > \boldsymbol{u}^T \boldsymbol{A}\boldsymbol{u}$. □

4.7.2 Proof of Lemma 4

This lemma stems from the fact that the first-order derivative of the function $f(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ is given by

$$\nabla f(\boldsymbol{x}) = \frac{1}{\|\boldsymbol{x}\|}\boldsymbol{I} - \frac{1}{\|\boldsymbol{x}\|^3}\boldsymbol{x}\boldsymbol{x}^T.$$

4.7.3 Proof of Theorem 4.1

The proof of Theorem 4.1 is given as follows.

[⇒] First, if Algorithm 4.1 converges locally uniformly to $\boldsymbol{x}_*$, our goal is to prove $\gamma < \lambda_{n-1}$. By contradiction, assume that $\gamma > \lambda_{n-1}$ [5]. Then choosing $\boldsymbol{x}^{(0)} = \boldsymbol{x}_* + \epsilon \boldsymbol{u}_{n-1}$, where $\boldsymbol{u}_{n-1}$ is the eigenvector corresponding to λ_{n-1}, leads to

$$\begin{aligned} & \boldsymbol{\delta}^{(1)} = \frac{|1-\alpha\lambda_{n-1}|}{1-\alpha\gamma}\boldsymbol{\delta}^{(0)} = \frac{1-\alpha\lambda_{n-1}}{1-\alpha\gamma}\boldsymbol{\delta}^{(0)} \\ \Rightarrow \quad & \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}_*\right\| = \left\|\boldsymbol{\delta}^{(1)}\right\| = \left|\frac{1-\alpha\lambda_{n-1}}{1-\alpha\gamma}\right| \|\epsilon\boldsymbol{u}_{n-1}\| = \frac{1-\alpha\lambda_{n-1}}{1-\alpha\gamma}\epsilon > \epsilon. \end{aligned}$$

This contradicts with the assumption that the sequence $\{\boldsymbol{x}^{(t)}\}$ lies inside the ϵ-vicinity of $\boldsymbol{x}_*$.

[⇐] Conversely, we will show that if $\boldsymbol{x}_*$ is a strict local minimum, then for any

[5] The case $\gamma = \lambda_{n-1}$ leads to convergence to a continuum which we leave as a future work.

$\alpha > 0$ such that $\alpha(\lambda_1 + \gamma) < 2$, Algorithm 4.1 with step size α converges locally uniformly to $\boldsymbol{x}_*$. By the same argument in [166], to prove the local stability of equation (4.4), it is sufficient to consider the linear equation without the quadratic residual

$$\boldsymbol{\delta}^{(t+1)} = \frac{1}{1-\alpha\gamma}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}(\boldsymbol{I} - \alpha\boldsymbol{A})\delta^{(t)}.$$

The above equation implies $\boldsymbol{\delta}^{(t)} = \boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{\delta}^{(t)}$ for $t = 1, 2, \ldots$. Thus, we have

$$\begin{aligned}\boldsymbol{\delta}^{(t+1)} &= \frac{\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}(\boldsymbol{I} - \alpha\boldsymbol{A})\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}}{1-\alpha\gamma}\delta^{(t)} = \frac{\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp} - \alpha\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}}{1-\alpha\gamma}\boldsymbol{\delta}^{(t)} \\ &= \frac{(\boldsymbol{I} - \alpha\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp})\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}}{1-\alpha\gamma}\boldsymbol{\delta}^{(t)} = \frac{\boldsymbol{I} - \alpha\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}}{1-\alpha\gamma}\boldsymbol{\delta}^{(t)}.\end{aligned}$$

Now consider the matrix $\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}$. There exists an eigenvalue decomposition $\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp}\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{x}_*}^{\perp} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{n-1}, \boldsymbol{x}_*]$ is an orthogonal matrix and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{n-1}, 0)$. Let $\boldsymbol{y}^{(t)} = \boldsymbol{U}^T\boldsymbol{\delta}^{(t)}$. Then

$$\boldsymbol{y}^{(t+1)} = \boldsymbol{U}^T\boldsymbol{\delta}^{(t+1)} = \frac{\boldsymbol{I} - \alpha\boldsymbol{\Lambda}}{1-\alpha\gamma}\boldsymbol{y}^{(t)} = \left(\frac{\boldsymbol{I} - \alpha\boldsymbol{\Lambda}}{1-\alpha\gamma}\right)^t\boldsymbol{y}^{(1)}. \tag{4.11}$$

In addition, since the last column of $\boldsymbol{U}$ is $\boldsymbol{x}_*$, we can compute the last element of $\boldsymbol{y}^{(1)}$ by

$$y_n^{(1)} = \boldsymbol{x}_*^T\boldsymbol{\delta}^{(1)} = \boldsymbol{x}_*^T(\boldsymbol{I} - \boldsymbol{x}_*\boldsymbol{x}_*^T)\boldsymbol{\delta}^{(0)} = 0. \tag{4.12}$$

From (4.11) and (4.12), we obtain

$$\left\|\boldsymbol{y}^{(t+1)}\right\| \leq \max_{1\leq i\leq n-1}\left|\frac{1-\alpha\lambda_i}{1-\alpha\gamma}\right|^t \cdot \left\|\boldsymbol{y}^{(1)}\right\|.$$

Since $\boldsymbol{x}_*$ is a strict local minimum, it follows from Lemma 4.2 that $\gamma < \lambda_{n-1} \leq \lambda_1$. Combining with the condition $\alpha\lambda_1 + \alpha\gamma < 2$, we obtain $\alpha\gamma < 1$. In order to show convergence, it remains to prove the inequality

$$\max_{1\leq i\leq n-1}\frac{|1-\alpha\lambda_i|}{1-\alpha\gamma} < 1 \Leftrightarrow |1-\alpha\lambda_i| < 1-\alpha\gamma,\ \forall\ 1\leq i\leq n-1.$$

Indeed, this inequality stems from the fact that $\alpha\lambda_1 + \alpha\gamma < 2$ and $\gamma < \lambda_{n-1} \leq \ldots \leq \lambda_1$.

4.7.4 Proof of Lemma 5

We have

$$\alpha_* = \underset{\substack{\alpha>0\\ \alpha(\lambda_1+\gamma)<2}}{\operatorname{argmin}} \max_{1\leq i\leq n-1}\frac{|1-\alpha\lambda_i|}{1-\alpha\gamma}. \tag{4.13}$$

For $\gamma < \lambda$, the function $\frac{1-\alpha\lambda}{1-\alpha\gamma}$ is monotonically decreasing. Denote $f(\alpha) = \max_{1\leq i\leq n-1}\frac{|1-\alpha\lambda_i|}{1-\alpha\gamma}$. Consider the following three cases:

- If $1-\alpha\lambda_{n-1} \geq 1-\alpha\lambda_1 \geq 0$, then (4.13) becomes

$$\min_{\alpha} f(\alpha) = \min_{\alpha\lambda_1 \leq 1} \frac{1-\alpha\lambda_{n-1}}{1-\alpha\gamma} = \begin{cases} f(\frac{1}{\lambda_1}) = \frac{\lambda_1-\lambda_{n-1}}{\lambda_1-\gamma} & \text{if } \lambda_1 > 0 \\ f(\infty) = \frac{\lambda_{n-1}}{\gamma} & \text{otherwise} \end{cases}$$

- If $1-\alpha\lambda_1 \leq 1-\alpha\lambda_{n-1} \leq 0$, then (4.13) becomes

$$\min_{\alpha} f(\alpha) = \min_{\alpha\lambda_{n-1} \geq 1} \frac{\alpha\lambda_1 - 1}{1-\alpha\gamma} = f\Big(\frac{1}{\lambda_{n-1}}\Big) = \frac{\lambda_1-\lambda_{n-1}}{\lambda_{n-1}-\gamma}.$$

- If $1-\alpha\lambda_1 \leq 0$ and $1-\alpha\lambda_{n-1} \geq 0$, then (4.13) becomes

$$\begin{aligned} \min_{\alpha} f(\alpha) &= \min_{\alpha(\lambda_1+\lambda_{n-1})\leq 2} \left\{ \frac{\alpha\lambda_1 - 1}{1-\alpha\gamma}, \frac{1-\alpha\lambda_{n-1}}{1-\alpha\gamma} \right\} \\ &= \begin{cases} f(\frac{2}{\lambda_1+\lambda_{n-1}}) = \frac{\lambda_1-\lambda_{n-1}}{\lambda_1+\lambda_{n-1}-2\gamma} & \text{if } \alpha(\lambda_1+\lambda_{n-1}) < 2 \\ f(\infty) = \frac{\lambda_{n-1}}{\gamma} & \text{otherwise} \end{cases} \end{aligned}$$

In summary, we have

- If $\lambda_1 + \lambda_{n-1} \leq 0$, then

$$\min_{\alpha} f(\alpha) = \min\Big\{ f\big(\frac{1}{\lambda_1}\big), f(\infty) \Big\} = f(\infty).$$

- If $\lambda_1 + \lambda_{n-1} > 0$, then

$$\min_{\alpha} f(\alpha) = \min\Big\{ f\big(\frac{1}{\lambda_1}\big), f\big(\frac{1}{\lambda_{n-1}}\big), f\big(\frac{2}{\lambda_1+\lambda_{n-1}}\big) \Big\} = f\big(\frac{2}{\lambda_1+\lambda_{n-1}}\big).$$

Chapter 5: On Local Linear Convergence of Projected Gradient Descent for Unit-Modulus Least Squares[1]

The unit-modulus least squares (UMLS) problem has a wide spectrum of applications in signal processing, e.g., phase-only beamforming, phase retrieval, radar code design, and sensor network localization. Scalable first-order methods such as projected gradient descent (PGD) have recently been studied as a simple yet efficient approach to solving the UMLS problem. Existing results on the convergence of PGD for UMLS often focus on global convergence to stationary points. As a non-convex problem, only a sublinear convergence rate has been established. However, these results do not explain the fast convergence of PGD frequently observed in practice. This chapter presents a novel analysis of convergence of PGD for UMLS, justifying the linear convergence behavior of the algorithm near the solution. By exploiting the local structure of the objective function and the constraint set, we establish an exact expression of the convergence rate and characterize the conditions for linear convergence. Simulations show that our theoretical analysis corroborates numerical examples. Furthermore, variants of PGD with adaptive step sizes are proposed based on the new insight revealed in our convergence analysis. The variants show substantial acceleration in practice.

[1]This work is currently under review and available at `https://arxiv.org/abs/2206.10832`.

5.1 Introduction

Unit-modulus least squares (UMLS) is formulated as the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{C}^N} \quad & \frac{1}{2}\|\boldsymbol{\Phi}\boldsymbol{w}-\boldsymbol{h}\|^2 \\ \text{s.t.} \quad & |w_i|^2 = 1 \text{ for } i = 1,\ldots,N, \end{aligned} \tag{5.1}$$

where $\boldsymbol{\Phi} \in \mathbb{C}^{M\times N}$ and $\boldsymbol{h} \in \mathbb{C}^M$. This problem arises in numerous machine learning and signal processing applications including, but not limited to, phase-only beamforming [138,206], phase retrieval [35,220], radar code design [184,203], and sensor network localization [70]. For instance, in phase-only beamforming applications, the goal is to design a weight vector $\boldsymbol{w}$ associated with N antennas so that it retains the power of each antenna and enhances reception of the signals from certain directions while mitigating interference located at other directions. For a uniform linear antenna array, $\boldsymbol{\Phi}$ can be the steering vector matrix with a Vandermonde structure.

It is well-known that UMLS is a non-convex NP-hard problem [237]. One traditional approach to this problem is semi-definite relaxation (SDR). In [141], Luo *et. al.* recast (5.1) as a quadratically constrained quadratic programming (QCQP) problem and then lifted it to an N^2-dimensional problem with a rank-1 constraint. By dropping the non-convex rank constraint, the resulting problem is convex and can be solved via interior point methods. The major disadvantage

of SDR is the high computational complexity ($O(N^7)$ flops and $O(N^2)$ memory), which is not suitable for large-scale problems in modern applications. Another approach that has recently been proposed by Tranter *et. al.* [206] is non-convex projected gradient descent (PGD). Since the projection onto the unit-modulus manifold is simple and low-cost, PGD was shown to be efficient in large-scale settings. Notably, the authors in [206] showed that, despite the lack of convexity, the algorithm converges globally to a set of stationary points of (5.1) and the rate of convergence is at least sublinear.

Motivated by Tranter's result, this chapter studies an in-depth convergence analysis of PGD for UMLS. First, we observe in practice that the algorithm frequently exhibits linear convergence near a local minimum of the problem. This is significantly faster than the sublinear convergence proven in [206]. Second, we believe the bound technique in [206] is rather conservative since it focuses on global characterization yet ignores the local structure of the problem around the solution. In particular, while UMLS is not a globally convex problem, it can still possess a benign geometry around a local minimum. In such scenario, one can expect that PGD will converge linearly to the local minimum similar to gradient descent for unconstrained minimization of a smooth and strongly convex function [160]. With this intuition, our goal here is to provide an analytical framework to uncover the fast linear convergence behavior of PGD near a local minimum of the UMLS problem.[2] By exploiting the local structure of the problem near local minima, we are

[2]Part of this work appeared in an earlier conference version [218], where we study the local convergence of PGD for minimizing a quadratic over the unit sphere. When $N = 1$, the UMLS

able to identify the sufficient conditions for local linear convergence of PGD with a fixed step size and obtain an exact expression of the convergence rate. In addition, we establish the region of convergence in which initializing the algorithm is guaranteed to converge to the desired local minimum. The theoretical rate predicts accurately the empirical convergence rate in our numerical simulation. Finally, in practical applications where prior knowledge of the solution is not available, we propose two adaptive-step-size variants of PGD that requires the same iteration complexity while offering faster linear convergence compared to the optimal fixed step size in theory.

The rest of the chapter is organized as follows. Section 5.2 presents the real-valued formulation of UMLS that is considered in this chapter and the PGD algorithm for solving this problem. Section 5.3 summarizes existing results on the convergence of PGD for UMLS in the literature, highlighting the fundamental similarity between the UMLS problem and the spherically constrained problem. Our convergence analysis is presented in Section 5.4, including solution properties, algorithm properties, and linear convergence properties. In Section 5.5, we propose two variants of PGD for UMLS that use adaptive step size schemes to effectively obtain fast linear convergence without the prior knowledge of the solution. Finally, in Section 5.6, we perform numerical experiments to verify our theoretical analysis.

problem and the spherically constrained LS problem coincide. For $N > 1$, UMLS introduces a more complex constraint set in the form of the cross product of multiple spherical constrains.

5.2 Problem Statement

In this section, we introduce fundamental concepts in formulating the UMLS problem as a standard constrained least-squares optimization and the PGD algorithm for solving it.

5.2.1 Notation

Throughout the chapter, we use the notations $\|\cdot\|_F$ and $\|\cdot\|_2$ to denote the Frobenius norm and the spectral norm of a matrix, respectively. Additionally, $\|\cdot\|$ is used on a vector to denote the Euclidean norm. Boldfaced symbols are reserved for vectors and matrices. The $t \times t$ identity matrix is denoted by $\boldsymbol{I}_t$. The t-dimensional vector of all zeros and the t-dimensional vector of all ones are denoted by $\mathbf{0}_t$ and $\mathbf{1}_t$, respectively. The notations $\otimes$ denotes the Kronecker product between two matrices and $\text{vec}(\cdot)$ denotes the vectorization of a matrix by stacking its columns on top of one another. For a complex number z, $\Re$ and $\Im$ denote the real and imaginary parts of z, respectively. Given an n-dimensional vector $\boldsymbol{x}$, x_i denotes the ith element of $\boldsymbol{x}$ and $\text{diag}(\boldsymbol{x})$ denotes the $n \times n$ diagonal matrix with the corresponding diagonal entries $x_1, \ldots, x_n$. Given a matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, the ith largest eigenvalue and the ith largest singular value of $\boldsymbol{X}$ are denoted by $\lambda_i(\boldsymbol{X})$ and $\sigma_i(\boldsymbol{X})$, respectively. The spectral radius of $\boldsymbol{X}$ is defined as $\rho(\boldsymbol{X}) = \max_i |\lambda_i(\boldsymbol{X})|$ and is less than or equal to the spectral norm, i.e., $\rho(\boldsymbol{X}) \leq \|\boldsymbol{X}\|_2$ [151]. If $\boldsymbol{X}$ is square and invertible, the condition number of $\boldsymbol{X}$ is defined as $\kappa(\boldsymbol{X}) = \sigma_1(\boldsymbol{X})/\sigma_n(\boldsymbol{X})$. Finally, we use $\boldsymbol{X} \succ 0$ to indicate the matrix $\boldsymbol{X}$ is positive definite (PD) and $\boldsymbol{X} \succeq 0$

to indicate the matrix $\boldsymbol{X}$ is positive semi-definite (PSD).

5.2.2 Real-valued Formulation of UMLS

For the convenience of analysis, we consider the following real-valued parametrization of (5.1):

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^{2N}} \quad & \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2 \\ \text{s.t.} \quad & x_{2i-1}^2 + x_{2i}^2 = 1 \text{ for } i = 1,\ldots,N, \end{aligned} \tag{5.2}$$

where $\boldsymbol{A} \in \mathbb{R}^{2M\times 2N}$ is partitioned into 2×2 blocks of form

$$\tilde{\boldsymbol{A}}_{ij} = \begin{bmatrix} \Re(\Phi_{ij}) & -\Im(\Phi_{ij}) \\ \Im(\Phi_{ij}) & \Re(\Phi_{ij}) \end{bmatrix}, \tag{5.3}$$

for $i = 1,\ldots,M$ and $j = 1,\ldots,N$. In addition, $\boldsymbol{x} = [\Re(w_1), \Im(w_1), \ldots, \Re(w_N), \Im(w_N)]^\top$ and $\boldsymbol{b} = [\Re(h_1), \Im(h_1), \ldots, \Re(h_M), \Im(h_M)]^\top$ are real-valued vectors. Next, we introduce the concepts of the 2-selection operator that selects the ith coordinate pair of a $2N$-dimensional vector. Since the unit-modulus constraint involves every pair of coordinates of $\boldsymbol{x}$, this operator allows us to simplify the representation of our result in the rest of the chapter:

Definition 5.1. *For each $i = 1,\ldots,N$, the ith **2-selection operator** is defined*

by $\boldsymbol{S}_i : \mathbb{R}^{2N} \to \mathbb{R}^2$ *such that*

$$\boldsymbol{S}_i(\boldsymbol{x}) = \begin{bmatrix} x_{2i-1} \\ x_{2i} \end{bmatrix},$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_{2N}]^\top$.

It is noteworthy that the 2-selection operators is linear. Using this operator, we can represent any vector $\boldsymbol{x} \in \mathbb{R}^{2N}$ as

$$\boldsymbol{x} = \sum_{i=1}^{N} \boldsymbol{e}_i \otimes \boldsymbol{S}_i(\boldsymbol{x}), \tag{5.4}$$

where $\boldsymbol{e}_i$ is the ith vector in the natural basis of $\mathbb{R}^N$. Then, we define the constraint set of the UMLS problem (5.6) based on the 2-selection operator.

Definition 5.2. *The* ***unit-modulus set*** *is defined by*

$$\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^{2N} : \|\boldsymbol{S}_i(\boldsymbol{x})\|^2 = 1, \forall i = 1, \ldots, N\}. \tag{5.5}$$

Using Definition 5.2, one can rewrite the optimization problem (5.2) as follows

$$\min_{\boldsymbol{x} \in \mathcal{C}} \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 . \tag{5.6}$$

For convenience, we denote the objective $f(\boldsymbol{x}) = \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$.

5.2.3 Projected Gradient Descent for UMLS

To define the projection onto the unit-modulus set $\mathcal{C}$, let us introduce the distance function from a point $\boldsymbol{x} \in \mathbb{R}^{2N}$ to $\mathcal{C}$ as

$$d(\boldsymbol{x}, \mathcal{C}) = \inf_{\boldsymbol{y} \in \mathcal{C}} \{\|\boldsymbol{y} - \boldsymbol{x}\|\}. \tag{5.7}$$

The set of all projections of $\boldsymbol{x}$ onto $\mathcal{C}$ is then given by

$$\Pi_{\mathcal{C}}(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathcal{C} \mid \|\boldsymbol{y} - \boldsymbol{x}\| = d(\boldsymbol{x}, \mathcal{C})\}. \tag{5.8}$$

It is well-known [210] that if $\mathcal{C}$ is closed, then for any $\boldsymbol{x} \in \mathbb{R}^n$, $\Pi_{\mathcal{C}}(\boldsymbol{x})$ is non-empty. Additionally, since the unit-modulus set $\mathcal{C}$ is non-convex, $\Pi_{\mathcal{C}}(\boldsymbol{x})$ can have more than one element. An orthogonal projection onto $\mathcal{C}$ is defined as $\mathcal{P}_{\mathcal{C}} : \mathbb{R}^{2N} \to \mathcal{C}$ such that $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ is chosen as an element of $\Pi_{\mathcal{C}}(\boldsymbol{x})$ based on a prescribed scheme (e.g., based on lexicographic order). In particular, we define the orthogonal projection $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ as projecting each coordinate pair of $\boldsymbol{x} \in \mathbb{R}^{2N}$ onto the unit 1-sphere

$$\boldsymbol{S}_i\big(\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})\big) = \begin{cases} \frac{\boldsymbol{S}_i(\boldsymbol{x})}{\|\boldsymbol{S}_i(\boldsymbol{x})\|} & \text{if } \boldsymbol{S}_i(\boldsymbol{x}) \neq \boldsymbol{0}_2, \\ [1, 0]^\top \triangleq \boldsymbol{s} & \text{if } \boldsymbol{S}_i(\boldsymbol{x}) = \boldsymbol{0}_2, \end{cases} \tag{5.9}$$

for $i = 1, \ldots, N$. Here we recall that $\boldsymbol{S}_i$ is defined in Definition 5.1. It is noted that when $\boldsymbol{S}_i(\boldsymbol{x}) = \boldsymbol{0}_2$, the set of projections of $\boldsymbol{0}_2$ onto the unit 1-sphere is non-singleton, i.e., the entire unit 1-sphere. In such case, we choose a certain element $\boldsymbol{s}$

Algorithm 5.1 Projected Gradient Descent (PGD)

Require: $\boldsymbol{x}^{(0)} \in \mathbb{R}^{2N}$
Ensure: $\{\boldsymbol{x}^{(k)}\}_{k=0}$
1: **for** $k = 0, 1, \ldots$ **do**
2: $\quad \boldsymbol{x}^{(k+1)} = \mathcal{P}_{\mathcal{C}}\big(\boldsymbol{x}^{(k)} - \eta \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b})\big)$
3: $\quad$ ▷ where $\mathcal{P}_{\mathcal{C}}$ is defined in (5.9)

in this set (e.g., $[1, 0]^\top$) as the value of $\boldsymbol{S}_i\big(\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})\big)$. We emphasize that this choice of the projection does not affect our subsequent analysis of local convergence.

Starting from some initial point $\boldsymbol{x}^{(0)}$, the PGD algorithm for solving (5.6) performs the following iterative update (see Algorithm 5.1):

$$\boldsymbol{x}^{(k+1)} = \mathcal{P}_{\mathcal{C}}\big(\boldsymbol{x}^{(k)} - \eta \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b})\big), \tag{5.10}$$

where $\eta > 0$ is a fixed step size. In the literature, PGD is also known as the gradient projection (GP) algorithm (e.g., [206]).

5.3 Preliminaries

This section presents a brief overview of existing results on convergence analysis of PGD for UMLS and the related problem of least squares with unit-norm constraint.

5.3.1 Existing Convergence Results on PGD for UMLS

The sublinear convergence of PGD to a set of stationary points of UMLS was studied in [206]. First, Tranter *et. al.* showed that any limiting point $\boldsymbol{x}^*$ of the

sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 5.1 is also a stationary point of (5.6). Second, they proved that for PGD with a fixed step size $0 < \eta < 1/\left\|\boldsymbol{A}\right\|_2^2$, the convergence of $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ to a set of stationary points of (5.6) is sublinear. In particular, the authors provided a sublinear bound on the distance between two consecutive iterates as follows[3]

$$\min_{0\leq l\leq k-1}\left\|\boldsymbol{x}^{(l+1)}-\boldsymbol{x}^{(l)}\right\| \leq \sqrt{\frac{2\eta\big(f(\boldsymbol{x}^{(0)})-f(\boldsymbol{x}^{*})\big)}{(1-\eta\left\|\boldsymbol{A}\right\|_2^2)k}}. \tag{5.11}$$

However, it is noted that the sublinear bound given by (5.11) is based on the worst-case analysis. In practice, we observe the algorithm enjoys fast linear convergence to a local minimum $\boldsymbol{x}^*$ of (5.6). Figure 5.1 illustrates the striking difference between the bound on $\left\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\right\|$ given by the RHS of (5.11) (sublinear in blue dashed line) and the corresponding empirical value obtained by running the PGD algorithm (linear in blue solid line). The additional bound on $\left\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\right\|$ (red dashed line) is derived from the bound on $\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}\right\|$ given later in (5.21) and the application of the triangle inequality: $\left\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\right\| \leq \left\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{*}\right\| + \left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}\right\|$. We observe the red dashed line and the blue solid line are parallel to each other, while the blue dashed line deviates quickly from the other two lines as k increases. In the next section, we study this unexplained convergence phenomenon of PGD for UMLS. We will provide exact formulations

[3]We note that in [206], the authors actually derived the convergence bound on a surrogate function $Q(\cdot)$ that quantifies the stationarity condition of (5.6). From Eqn. (23b) in [206], we have the value of $Q(\cdot)$ at iteration k equals to $\frac{1}{\eta^2}\left\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\right\|^2$. In the literature, such convergence metric is related to the generalized gradient norm, (e.g., [12]-Section 2.3.2).

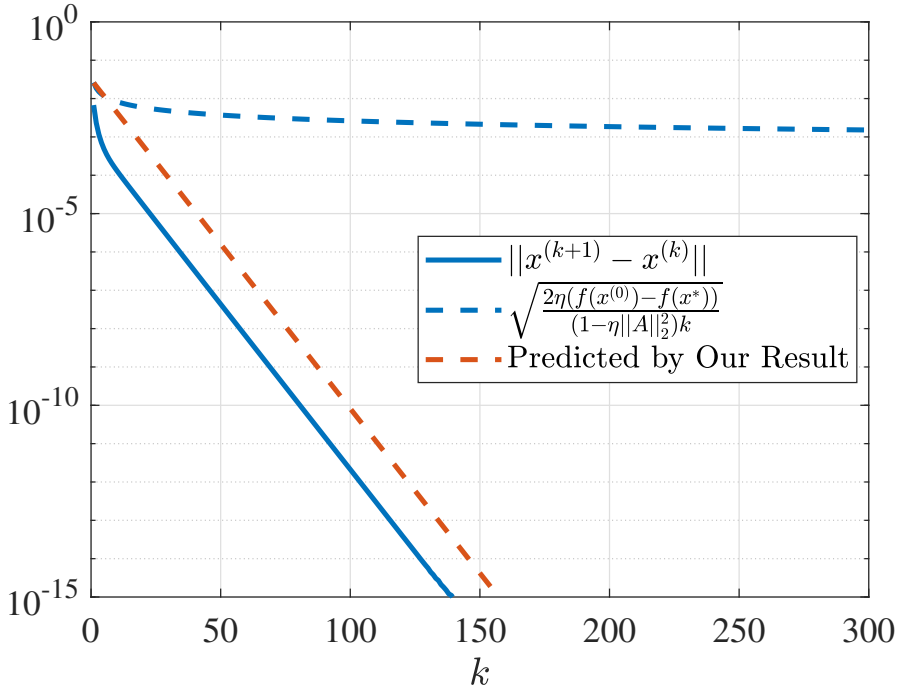


Figure 5.1: Plot of $\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right\|$ (blue solid) generated by PGD for UMLS with a fixed step size $\eta = 0.9/\left\|\boldsymbol{A}\right\|_2^2$. The blue dashed line represents the sublinear bound given by (5.11). The red dashed line is based on our linear upper bound proposed in this work. Further details of the data generated for this figure are given later in our simulation in Section 5.6.

of the region of convergence and the linear convergence rate. The selection of the fixed step size $\eta < 1/\left\|\boldsymbol{A}\right\|_2^2$ is conservative as it may not be the optimal choice to attain a quick convergence speed. We will demonstrate in our simulation that larger step sizes enable faster convergence of PGD for UMLS.

5.3.2 Least Squares with Unit-Norm Constraint

A closely-related problem to UMLS is the unit-norm least squares (UNLS)

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^N} \quad & \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\right\|^2 \\ \text{s.t.} \quad & \left\|\boldsymbol{x}\right\|^2 = 1, \end{aligned} \tag{5.12}$$

where $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{b} \in \mathbb{R}^N$. While UMLS requires each of the N coordinate pairs of the solution lies on the unit 1-sphere, UNLS requires the solution itself lies on the $N-1$-sphere. Unlike the case of unit-modulus constraint, minimizing a quadratic form over the unit sphere is not NP-hard and is solvable as an eigenvalue problem [87, 187]. The convergence of PGD for UNLS has recently been studied in [13, 218]. Table 5.1 summarizes the existing convergence result on UNLS and the new convergence result on UMLS we derive in this chapter, highlighting the connection between the two works.

5.4 Convergence Analysis

This section presents the convergence analysis of PGD for UMLS. We begin with the properties of the solution of the problem and the PGD algorithm. Next, we present the main result on the convergence of PGD for UMLS. Finally, we provide the detailed proof at the end of the section.

[4]This is a more intuitive but not the most general constraint on the step size. The original version of this condition on the step size is given in Theorem 5.1.

Table 5.1: Comparison between the existing convergence analysis of PGD for least squares with unit-norm constraint [218] and the novel convergence analysis of PGD for unit-modulus constraint proposed in this work. In each case, $\boldsymbol{x}^*$ is a stationary point and $\boldsymbol{Z}$ is a basis matrix for the null space of the Jacobian of all constraints at $\boldsymbol{x}^*$.

	Unit-norm constraint [218]	**Unit-modulus constraint** (this work)
Problem formulation	$\min_{\boldsymbol{x}\in\mathbb{R}^N} \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\right\|^2$ s.t. $\left\|\boldsymbol{x}\right\|=1$	$\min_{\boldsymbol{x}\in\mathbb{R}^{2N}} \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\right\|^2$ s.t. $\left\|\boldsymbol{S}_i(\boldsymbol{x})\right\|=1, \forall i=1,\ldots,N$
First-order necessary condition	$\exists\gamma\in\mathbb{R}: \boldsymbol{A}^\top(\boldsymbol{A}^\top\boldsymbol{x}^*-\boldsymbol{b})=\gamma\boldsymbol{x}^*$	$\exists\boldsymbol{\gamma}\in\mathbb{R}^N: \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^*-\boldsymbol{b})=(\mathrm{diag}(\boldsymbol{\gamma})\otimes\boldsymbol{I}_2)\boldsymbol{x}^*$
Reduced Riemannian Hessian	$\boldsymbol{H}=\boldsymbol{Z}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{Z}-\gamma\boldsymbol{I}_N$	$\boldsymbol{H}=\boldsymbol{Z}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{Z}-\mathrm{diag}(\boldsymbol{\gamma})$
Second-order *necessary* condition	$\boldsymbol{H}\succeq\boldsymbol{0}_N$	$\boldsymbol{H}\succeq\boldsymbol{0}_N$
Second-order *sufficient* condition	$\boldsymbol{H}\succ\boldsymbol{0}_N$	$\boldsymbol{H}\succ\boldsymbol{0}_N$
Fixed-point condition on step size	$1-\eta\gamma>0$	$\boldsymbol{I}_N-\eta\,\mathrm{diag}(\boldsymbol{\gamma})\succ\boldsymbol{0}_N$
Convergence condition on the step size	$\eta\big(\lambda_1(\boldsymbol{H})+2\gamma\big)<2$	$\eta\big(\lambda_1(\boldsymbol{H})+2\max_i\gamma_i\big)<2$ [4]
Linear convergence rate	$\rho\big(\boldsymbol{I}_N-\eta(1-\eta\gamma)^{-1}\boldsymbol{H}\big)$	$\rho\big(\boldsymbol{I}_N-\eta(\boldsymbol{I}_N-\eta\,\mathrm{diag}(\boldsymbol{\gamma}))^{-1}\boldsymbol{H}\big)$

5.4.1 Solution Properties

The Lagrange function corresponding to (5.6) is given by

$$L(\boldsymbol{x}, \boldsymbol{\gamma}) = \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 - \frac{1}{2} \sum_{i=1}^{N} \gamma_i (x_{2i-1}^2 + x_{2i}^2 - 1),$$

where $\boldsymbol{\gamma} \in \mathbb{R}^N$ is the Lagrange multiplier. The derivatives of L with respect to $\boldsymbol{x}$ can be computed as

$$\begin{cases} \nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\gamma}) &= \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) - (\mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2)\boldsymbol{x}, \\ \nabla_{\boldsymbol{x}}^2 L(\boldsymbol{x}, \boldsymbol{\gamma}) &= \boldsymbol{A}^\top \boldsymbol{A} - \mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2. \end{cases} \tag{5.13}$$

It can be shown that any feasible point $\boldsymbol{x} \in \mathcal{C}$ is also a regular point of the constraint set. Specifically, we first represent the constraints as $\boldsymbol{h} : \mathbb{R}^{2N} \to \mathbb{R}^N$ such that $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}_N$, where $h_i(\boldsymbol{x}) = \|\boldsymbol{S}_i(\boldsymbol{x})\|^2 - 1$ for $i = 1, \ldots, N$. Then, the Jacobian of all the constraints at $\boldsymbol{x}$, defined as $J_{ij} = \partial h_i(\boldsymbol{x}) / \partial x_j$, is given by

$$\boldsymbol{J}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{e}_1^\top \otimes \boldsymbol{S}_1^\top(\boldsymbol{x}) \\ \cdots \\ \boldsymbol{e}_N^\top \otimes \boldsymbol{S}_N^\top(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^{N \times 2N}.$$

Since $\boldsymbol{J}(\boldsymbol{x})$ is full row rank for any $\boldsymbol{x} \in \mathcal{C}$, we have $\boldsymbol{x}$ is a regular point of the constraint set (see Chapter 11 in [140]). The following lemma establishes the first-order necessary conditions for local optima of UMLS problems.

Lemma 5.1. *The first-order necessary conditions for $\boldsymbol{x}^* \in \mathbb{R}^{2N}$ to be a local*

minimum of (5.6) are $\boldsymbol{x}^* \in \mathcal{C}$ *and there exists a Lagrange multiplier* $\boldsymbol{\gamma} \triangleq \boldsymbol{\gamma}(\boldsymbol{x}^*) \in \mathbb{R}^N$ *such that*

$$\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) = (\mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2)\boldsymbol{x}^*. \tag{5.14}$$

Any point satisfying the foregoing first-order necessary conditions is called a ***stationary*** *point of (5.6).*

By setting $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\gamma})$ in (5.13) to $\mathbf{0}$, the proof of Lemma 5.1 follows the same derivation in [140]-Chapter 11.3. Next, we examine the second-order conditions for local optima of problem (5.6) via the basis of the tangent plane to $\mathcal{C}$ at $\boldsymbol{x}^*$. The following lemma provides further insight into these conditions.

Lemma 5.2. *Let* $\boldsymbol{x}^*$ *be a stationary point of problem (5.6) with the corresponding Lagrange multiplier* $\boldsymbol{\gamma}$*. A basis of the tangent space to* $\mathcal{C}$ *at* $\boldsymbol{x}^*$ *is given by the semi-orthogonal matrix* $\boldsymbol{Z} \in \mathbb{R}^{2N \times N}$ *such that*

$$\boldsymbol{Z} = \sum_{i=1}^{N} \boldsymbol{e}_i \boldsymbol{e}_i^\top \otimes \boldsymbol{v}_i, \tag{5.15}$$

where $\boldsymbol{v}_i = [-x^*_{2i}, x^*_{2i-1}]^\top$*. Denote the* ***reduced Riemannian Hessian*** *associated with* $\boldsymbol{x}^*$ *by*

$$\boldsymbol{H} = \boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{Z} - \mathrm{diag}(\boldsymbol{\gamma}). \tag{5.16}$$

The second-order necessary condition for $\boldsymbol{x}^*$ *to be a local minimum of (5.6) is*

$\boldsymbol{H} \succeq \mathbf{0}_N$. *The second-order sufficient condition for* $\boldsymbol{x}^*$ *to be a* ***strict*** *local minimum of (5.6) is* $\boldsymbol{H} \succ \mathbf{0}_N$.

The proof of Lemma 5.2 is given in Appendix 5.8.1.

Remark 5.1. *The concept of Riemannian Hessian has been well-studied in differential geometry (e.g., [124]). From (5.16), one can see that the first term takes into account the curvature of the objective function restricted to the unit-modulus manifold* $\mathcal{C}$. *On the other hand, the second term characterizes the curvature of the manifold* $\mathcal{C}$. *While this is an elementary result in differential geometry, we include the proof detail in Appendix 5.8.2 for self-containedness.*

5.4.2 Algorithm Properties

The PGD algorithm can be viewed as a fixed-point iteration and hence, can be analyzed via the existing tools from fixed-point theory. We first define the convergent point of the PGD update (5.10) as follows.

Definition 5.3. *The point* $\boldsymbol{x} \in \mathcal{C}$ *is a* ***fixed point*** *of Algorithm 5.1 with step size* $\eta > 0$ *if it satisfies*

$$\boldsymbol{x} = \mathcal{P}_{\mathcal{C}}\big(\boldsymbol{x} - \eta \boldsymbol{A}^{\top}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\big). \tag{5.17}$$

If the constraint set $\mathcal{C}$ is convex, any fixed point of Algorithm 5.1 is also an optimal solution of the constrained least squares problem [16]. Since the unit-modulus

constraint set is non-convex, we show that any fixed point of Algorithm 5.1 is a stationary point of (5.6) as follows.

Lemma 5.3. *The vector $\boldsymbol{x}^*$ is a fixed point of Algorithm 5.1 with step size $\eta > 0$ if and only if $\boldsymbol{x}^*$ is a stationary point of the non-convex problem (5.6) and the corresponding Lagrange multiplier $\boldsymbol{\gamma}$ satisfies*

$$\begin{cases} \gamma_i < 1/\eta & \text{if } \boldsymbol{S}_i(\boldsymbol{x}^*) \neq \boldsymbol{s} \\ \gamma_i \leq 1/\eta & \text{if } \boldsymbol{S}_i(\boldsymbol{x}^*) = \boldsymbol{s} \end{cases} \quad \forall i = 1, \ldots, N, \tag{5.18}$$

where $\boldsymbol{s}$ is defined in (5.9).

The proof of this lemma is given in Appendix 5.8.3. Lemma 5.3 suggests that when η is sufficiently small, all stationary points of (5.6) can be fixed points of Algorithm 5.1. As the step size η increases, only fewer stationary points satisfying (5.18) can be fixed points of the algorithm. Next, we study the first-order Taylor expansion of the projection $\mathcal{P}_{\mathcal{C}}$ about a point in $\mathcal{C}$ in the following proposition:

Proposition 5.1. *For any $\boldsymbol{x} \in \mathcal{C}$ and $\boldsymbol{\delta} \in \mathbb{R}^{2N}$, we have*

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x} + \boldsymbol{\delta}) = \boldsymbol{x} + \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{\delta} + \boldsymbol{q}(\boldsymbol{\delta}), \tag{5.19}$$

where $\boldsymbol{Z} = \sum_{i=1}^{N} \boldsymbol{e}_i \boldsymbol{e}_i^\top \otimes \boldsymbol{v}_i$, for $\boldsymbol{v}_i = [-x_{2i}, x_{2i-1}]^\top$, and $\boldsymbol{q} : \mathbb{R}^{2N} \to \mathbb{R}^{2N}$ satisfies $\|\boldsymbol{q}(\boldsymbol{\delta})\| \leq 2\|\boldsymbol{\delta}\|^2$.

The proof of this proposition is given in Appendix 5.8.4. It is noteworthy from Proposition 5.1 that the projection $\mathcal{P}_{\mathcal{C}}$ is differentiable at any $\boldsymbol{x} \in \mathcal{C}$. Second, the

derivative of $\mathcal{P}_{\mathcal{C}}$, given by $\boldsymbol{Z}\boldsymbol{Z}^\top$, coincides with the orthogonal projection onto the tangent space of $\mathcal{C}$ at $\boldsymbol{x}$ [129]. Third, the expansion (5.19) is universal, regardless of the magnitude of $\boldsymbol{\delta}$.

5.4.3 Main Result

We are now in position to state our main result on the linear convergence of PGD for UMLS.

Theorem 5.1. *Consider a stationary point $\boldsymbol{x}^* \in \mathcal{C}$ of the UMLS problem (5.6) with the corresponding Lagrange multiplier $\boldsymbol{\gamma} \triangleq \boldsymbol{\gamma}(\boldsymbol{x}^*) \in \mathbb{R}^N$ defined in (5.14) and the reduced Riemannian Hessian $\boldsymbol{H} \triangleq \boldsymbol{H}(\boldsymbol{x}^*) \in \mathbb{R}^{N\times N}$ defined in (5.16). Let $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty} \subset \mathbb{R}^{2N}$ be the sequence generated by Algorithm 5.1 with a fixed step size $\eta > 0$. Assume that*

(C1) $\boldsymbol{H} \succ \boldsymbol{0}_N$ (sufficient condition for $\boldsymbol{x}^*$ being a strict local minimum),

(C2) $\eta\gamma_i \neq 1$ for all $i = 1, \ldots, N$, and

(C3) $\rho(\boldsymbol{M}_\eta) < 1$ where

$$\boldsymbol{M}_\eta = \boldsymbol{I}_N - \eta\Big(\boldsymbol{I}_N - \eta\,\mathrm{diag}(\boldsymbol{\gamma})\Big)^{-1}\boldsymbol{H}. \tag{5.20}$$

Then, there exists a finite constant $c_0(\boldsymbol{x}^*, \eta)$ [5] such that for any $\boldsymbol{x}^{(0)} \in \mathcal{C}$ satisfying $\big\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\big\| < c_0(\boldsymbol{x}^*, \eta)$, the sequence $\{\big\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\big\|\}_{k=0}^{\infty}$ converges to 0. Further-

[5] A closed-form expression of $c_0(\boldsymbol{x}^*, \eta)$ is given in Lemma 5.9.

more, if $\left\|\boldsymbol{x}^{(0)}-\boldsymbol{x}^*\right\| < \rho(\boldsymbol{M}_\eta)c_0(\boldsymbol{x}^*,\eta)$*, it holds for any integer* $k \geq 0$ *that*

$$\frac{\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\|}{\left\|\boldsymbol{x}^{(0)}-\boldsymbol{x}^*\right\|} < \left(1-\frac{\left\|\boldsymbol{x}^{(0)}-\boldsymbol{x}^*\right\|}{\rho(\boldsymbol{M}_\eta)c_0(\boldsymbol{x}^*,\eta)}\right)^{-1}\rho^k(\boldsymbol{M}_\eta). \tag{5.21}$$

In (5.21), Algorithm 5.1 with fixed step size η *is said to converge* ***linearly*** *to* $\boldsymbol{x}^*$ *with a rate of* $\rho(\boldsymbol{M}_\eta)$*.*

Theorem 5.1 suggests that PGD in Algorithm 5.1 initialized near a strict local minimum as indicated by (C1) with a proper step size η following the requirements in (C2) and (C3) converges linearly to the local minimum. The theorem establishes three key results for the linear convergence of Algorithm 5.1: the region of convergence, the rate of convergence, and the bound on the error through iterations. Notably, while the previous result in [206] proves the sublinear convergence to a set of stationary points of (5.6), our result in Theorem 5.1 shows the linear convergence to a strict local minimum. It is worthwhile mentioning that the linear convergence of $\{\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\|\}_{k=0}^{\infty}$ given by (5.21) matches with the definition of R-linear convergence in [112]-Appendix A.[6]

Note that Theorem 5.1 does not explicitly suggest an upper bound on η that ensures convergence and it may appear that PGD with arbitrarily large step size η still converges. However, to ensure convergence, the implicit condition on η in (C3)

[6]Compared to Q-linear convergence, R-linear convergence concerns the overall rate of decrease in the error, rather than the decrease over each individual step of the algorithm. A more elaborate bound on the convergence of non-linear difference equations of form (5.61) is developed in [216], in terms of the number of iterations to reach certain accuracy. In this work, we use a simpler result in Lemmas 5.13 and 5.14 to demonstrate the linear convergence.

must hold. To provide an intuition for the step size requirement in this condition, let us consider a more restrictive condition that suffices (C3):

Lemma 5.4. *Let $\eta > 0$ be a step size such that*

(C3') $\eta(\lambda_1(\boldsymbol{H}) + 2\overline{\gamma}) < 2$, *where* $\overline{\gamma} = \max_i \gamma_i$.

Then, Condition (C3) in Theorem 5.1 holds, i.e., $\rho(\boldsymbol{M}_\eta) < 1$.

The proof of Lemma 5.4 is given in Appendix 5.8.5. When $\lambda_1(\boldsymbol{H}) + 2\overline{\gamma} \leq 0$, any step size $\eta > 0$ satisfies (C3') and hence, satisfies (C3). When $\lambda_1(\boldsymbol{H}) + 2\overline{\gamma} > 0$, (C3') suggests an upper bound on η that is sufficient but not necessary for (C3), i.e., $\eta < 2/(\lambda_1(\boldsymbol{H}) + 2\overline{\gamma})$. As can be seen from Table 5.1, Condition (C3') is similar to the convergence condition in the case of unit-norm constraint.

In Theorem 5.1, Condition (C3) suggest a non-linear relationship between the convergence rate $\rho(\boldsymbol{M}_\eta)$ and the step size η. In principle, one can find the optimal step size for local linear convergence by solving the 1-D optimization

$$\begin{aligned}\eta^* &= \operatorname*{argmin}_{\eta>0} \rho\big(\boldsymbol{M}_\eta(\boldsymbol{x}^*)\big) \\ &= \operatorname*{argmin}_{\eta>0} \rho\Big(\boldsymbol{I}_N - \eta\big(\boldsymbol{I}_N - \eta \operatorname{diag}(\boldsymbol{\gamma}(\boldsymbol{x}^*))\big)^{-1}\boldsymbol{H}(\boldsymbol{x}^*)\Big).\end{aligned}$$

In the last equation, we spell out the dependence on $\boldsymbol{x}^*$ to emphasize that the prior knowledge of the local minimum is critical for determining the optimal step size. In the next section, we propose two variants of PGD with adaptive step size schemes that do not require prior knowledge of $\boldsymbol{M}_\eta$ to select the optimal step size.

The proposed algorithms enjoy the fast convergence of PGD with a fixed optimal step size while remaining the same computational complexity per iteration.

5.4.4 Proof of Theorem 5.1

This subsection presents a proof of Theorem 5.1, arranging the key ideas into lemmas and deferring their proofs to the appendix. Let us begin with the claim that the strict local minimum $\boldsymbol{x}^*$ in Theorem 5.1 is also a fixed point of PGD with the appropriate choice of the step size η.

Lemma 5.5. *Consider the same setting as Theorem 5.1. Assume that Conditions (C1)-(C3) in Theorem 5.1 hold. Then, $\boldsymbol{x}^*$ is a fixed point of Algorithm 5.1 with the given step size η and its corresponding Lagrange multiplier $\boldsymbol{\gamma}$ satisfies $\gamma_i < 1/\eta$, for all $i = 1, \ldots, N$.*

The proof of Lemma 5.5 is given in Appendix 5.8.6. Next, we establish a recursion on the error vector, based on the first-order approximation of the projection in Proposition 5.1.

Lemma 5.6. *Consider the same setting as Theorem 5.1. Assume that Conditions (C1)-(C3) in Theorem 5.1 hold. Let $\boldsymbol{D}_\eta = (\boldsymbol{I}_N - \eta \operatorname{diag}(\boldsymbol{\gamma}))^{-1}$ and $\boldsymbol{\delta}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^*$ be the error vector at the kth iteration of Algorithm 5.1. Then, for any integer*

$k \geq 0$, we have

$$\boldsymbol{\delta}^{(k+1)} = \boldsymbol{Z}\boldsymbol{Z}^\top(\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)} + \boldsymbol{q}\big((\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}\big), \tag{5.22}$$

where $\boldsymbol{Z}$ at $\boldsymbol{x}^*$ and $\boldsymbol{q}$ are defined in Proposition 5.1.

The proof of Lemma 5.6 is given in Appendix 5.8.7. Equation (5.22) can be viewed as an approximately linear dynamic on the error $\boldsymbol{\delta}^{(k)}$. As the error becomes sufficiently small, the residual term $\boldsymbol{q}((\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)})$ is negligible while the linear term $\boldsymbol{Z}\boldsymbol{Z}^\top(\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}$ dominates. It has been well-studied in the literature [15, 166, 215, 216, 218] that the linear convergence rate of (5.22) is the spectral radius of the linear operator $\boldsymbol{Z}\boldsymbol{Z}^\top(\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})$. However, following the argument about the structural constraint on the error vector in [215], we emphasize the fact $\boldsymbol{\delta}^{(k)} = \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* + \boldsymbol{\delta}^{(k)}) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)$ is the difference between two points on the unit-modulus manifold and show that the error vector is dominated by the component on the tangent plane to $\mathcal{C}$ at $\boldsymbol{x}^*$.

Lemma 5.7. *Consider the same setting as Theorem 5.1. At the kth iteration of Algorithm 5.1, we have*

$$\boldsymbol{\delta}^{(k)} = \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{\delta}^{(k)} + \boldsymbol{q}(\boldsymbol{\delta}^{(k)}), \tag{5.23}$$

where $\boldsymbol{Z}$ at $\boldsymbol{x}^*$ and $\boldsymbol{q}$ are defined in Proposition 5.1.

The proof of Lemma 5.7 is given in Appendix 5.8.8. Next, combining Lemmas 5.6

and 5.7, we obtain a recursion on the error vector that implicitly enforces it to lie on the tangent plane to $\mathcal{C}$ at $\boldsymbol{x}^*$ as follows.

Lemma 5.8. *Consider the same setting as Theorem 5.1. Assume that Conditions (C1)-(C3) in Theorem 5.1 hold. Then by Lemmas 5.6 and 5.7, the error vector at the kth iteration of Algorithm 5.1 satisfies*

$$\boldsymbol{\delta}^{(k+1)} = \boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top\boldsymbol{\delta}^{(k)} + \hat{\boldsymbol{q}}(\boldsymbol{\delta}^{(k)}), \tag{5.24}$$

where $\boldsymbol{Z}$ at $\boldsymbol{x}^*$ is defined in Proposition 5.1, $\hat{\boldsymbol{q}} : \mathbb{R}^{2N} \to \mathbb{R}^{2N}$ satisfies $\|\hat{\boldsymbol{q}}(\boldsymbol{\delta})\| \leq 2c_\eta(c_\eta+1)\|\boldsymbol{\delta}\|^2$, and $c_\eta = \left\|((\boldsymbol{I}_N - \eta\,\mathrm{diag}(\boldsymbol{\gamma}))^{-1} \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\right\|_2$.

The proof of Lemma 5.8 is given in Appendix 5.8.9. Finally, we show that the convergence of the sequence $\{\boldsymbol{\delta}^{(k)}\}_{k=0}^{\infty}$ by recognizing that *(i)* the spectral radius of $\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top$ is the same as that of $\boldsymbol{M}_\eta$ and *(ii)* the recursion (5.24) is an approximately linear difference equation that is convergent for $\boldsymbol{\delta}^{(0)}$ sufficiently close to $\boldsymbol{0}_{2N}$.

Lemma 5.9. *Consider the same setting as Theorem 5.1. Assume that Conditions (C1)-(C3) in Theorem 5.1 hold. Let us define $\overline{\gamma} = \max_i \gamma_i$, $\underline{\gamma} = \min_i \gamma_i$ and*

$$c_0(\boldsymbol{x}^*, \eta) = \frac{1-\rho(\boldsymbol{M}_\eta)}{2c_\eta(c_\eta+1)}\frac{1-\eta\overline{\gamma}}{1-\eta\underline{\gamma}}, \tag{5.25}$$

where c_η is defined in Lemma 5.8. If $\left\|\boldsymbol{\delta}^{(0)}\right\| < c_0(\boldsymbol{x}^*, \eta)$, then the sequence $\{\boldsymbol{\delta}^{(k)}\}_{k=0}^{\infty}$ converges to $\boldsymbol{0}_{2N}$. Furthermore, let $c_1(\boldsymbol{x}^*, \eta) = \rho(\boldsymbol{M}_\eta)c_0(\boldsymbol{x}^*, \eta)$. Then, for any

$\left\|\boldsymbol{\delta}^{(0)}\right\| < c_1(\boldsymbol{x}^*, \eta)$ *and integer* $k \geq 0$*, we have*

$$\left\|\boldsymbol{\delta}^{(k)}\right\| \leq \left(1 - \frac{\left\|\boldsymbol{\delta}^{(0)}\right\|}{c_1(\boldsymbol{x}^*, \eta)}\right)^{-1} \left(\frac{1 - \eta\underline{\gamma}}{1 - \eta\overline{\gamma}}\right)^{1/2} \left\|\boldsymbol{\delta}^{(0)}\right\| \rho^k(\boldsymbol{M}_\eta). \tag{5.26}$$

The proof of Lemma 5.9 is given in Appendix 5.8.10. With this lemma, we complete our proof of Theorem 5.1.

5.5 Implementation Aspects

This subsection describes two practical variants of PGD with adaptive step size that can be used when no prior knowledge of the solution is available: PGD with backtracking line search (Algorithm 5.2) and Nesterov's accelerated PGD with adaptive restart (Algorithm 5.3).

5.5.1 Backtracking PGD (Bt-PGD)

In backtracking PGD, the step size is chosen to approximately minimize the objective function $f(\boldsymbol{x}) = \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\right\|^2$ along the ray $\{\boldsymbol{x} - \eta\tilde{\boldsymbol{g}}_\eta \mid \eta > 0\}$, where

$$\tilde{\boldsymbol{g}}_\eta = \frac{1}{\eta}\Big(\boldsymbol{x} - \mathcal{P}_{\mathcal{C}}\big(\boldsymbol{x} - \eta\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\big)\Big)$$

Algorithm 5.2 Backtracking PGD (Bt-PGD)

Require: $\boldsymbol{x}^{(0)} \in \mathbb{R}^{2N}$, $\alpha \in (0,1]$, $\beta \in (0,1)$
Ensure: $\{\boldsymbol{x}^{(k)}\}_{k=0}$
1: $\eta_0 = 1$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad \boldsymbol{g}_k = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b})$
4: $\quad \eta_k = \eta_k/\beta$
5: $\quad$ **repeat**
6: $\quad\quad \eta_k = \beta\eta_k$
7: $\quad\quad \tilde{\boldsymbol{g}}_{\eta_k} = \frac{1}{\eta_k}\big(\boldsymbol{x}^{(k)} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^{(k)} - \eta_k \boldsymbol{g}_k)\big)$
8: $\quad$ **until** $\tilde{\boldsymbol{g}}_{\eta_k}^\top \boldsymbol{A}^\top \boldsymbol{A}\tilde{\boldsymbol{g}}_{\eta_k} \leq \frac{1}{\eta_k}\left\|\tilde{\boldsymbol{g}}_{\eta_k}\right\|^2$
9: $\quad \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \eta_k \tilde{\boldsymbol{g}}_{\eta_k}$
10: $\quad \eta_{k+1} = \eta_k/\alpha$

is the generalized gradient. To guarantee certain decrease in the objective function, we use the following backtracking condition [12]

$$f(\boldsymbol{x} - \eta\tilde{\boldsymbol{g}}_\eta) \leq f(\boldsymbol{x}) - \eta\tilde{\boldsymbol{g}}_\eta^\top \nabla f(\boldsymbol{x}) + \frac{\eta}{2}\left\|\tilde{\boldsymbol{g}}_\eta\right\|^2. \tag{5.27}$$

Since $f(\cdot)$ is a quadratic, it can be expanded as

$$f(\boldsymbol{x} - \eta\tilde{\boldsymbol{g}}_\eta) = f(\boldsymbol{x}) - \eta\tilde{\boldsymbol{g}}_\eta^\top \nabla f(\boldsymbol{x}) + \eta^2 \tilde{\boldsymbol{g}}_{\eta_k}^\top \nabla^2 f \tilde{\boldsymbol{g}}_{\eta_k}. \tag{5.28}$$

Substituting (5.28) back into the LHS of (5.27) and using the fact that $\nabla^2 f = \boldsymbol{A}^\top\boldsymbol{A}$, we obtain the simplified backtracking condition $\tilde{\boldsymbol{g}}_{\eta_k}^\top \boldsymbol{A}^\top \boldsymbol{A}\tilde{\boldsymbol{g}}_{\eta_k} \leq \frac{1}{\eta_k}\left\|\tilde{\boldsymbol{g}}_{\eta_k}\right\|^2$ as in Algorithm 5.2-Line 8. It is worthwhile to note that a factor of $1/\alpha$ is applied to increase the step size at the end of each iteration to encourage the algorithm to explore larger step sizes with faster convergence. We emphasize that this strategy

is different from the well-known backtracking line search method in the literature (e.g., [24]), in which the step size η is reset to 1 before the backtracking line search is performed. As a result, the constant α in Algorithm 5.2 should not be interpreted as the fraction of the decrease in the objective function as in [24]-Algorithm 9.2.

5.5.2 Adaptive Restart Nesterov's Accelerated PGD (ARNAPGD)

Next, we present an acceleration technique for PGD, named adaptive restart Nesterov's accelerated projected gradient descent (ARNAPGD). In unconstrained optimization, it has been well-known that Nesterov's accelerated gradient (NAG) [160] can dramatically improve the linear convergence rate of gradient descent (GD) for minimizing a μ-strongly convex, L-smooth function. As pointed out in [128]-Proposition 12, GD with a fixed step size $\alpha = 1/L$ has convergence rate $\rho \leq \sqrt{(L-\mu)/(L+\mu)}$, while NAG with fixed parameters $\alpha = 1/L$ and $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$ has convergence rate $\rho \leq \sqrt{1 - \sqrt{\mu/L}}$. Since NAG requires a specific choice of parameters that depends on L and μ, Donoghue and Candes [164] proposed a more practical variant called the Nesterov's accelerated gradient with adaptive restart (ARNAG) that recovers the same rate of convergence with no prior knowledge of function parameters. In this work, we modify ARNAG with gradient scheme to the context of PGD for constrained optimization. Specifically, each iteration uses backtracking line search for determining the *projected* gradient step η and the *generalized* gradient scheme for determining when to restart the momentum. The advantage of this acceleration is it has the same

Algorithm 5.3 Adaptive restart Nesterov's accelerated PGD (ARNAPGD) with gradient scheme

Require: $\boldsymbol{x}^{(0)} \in \mathbb{R}^{2N}$, $\alpha \in (0, 1]$, $\beta \in (0, 1)$
Ensure: $\{\boldsymbol{x}^{(k)}\}_{k=0}$
$\eta_0 = 1$
$\theta_0 = 1$
$\boldsymbol{y}^{(0)} = \boldsymbol{x}^{(0)}$
for $k = 0, 1, 2, \ldots$ **do**
    $\boldsymbol{g}_k = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{y}^{(k)} - \boldsymbol{b})$
    $\eta_k = \eta_k/\beta$
    repeat
        $\eta_k = \beta\eta_k$
        $\tilde{\boldsymbol{g}}_{\eta_k} = \frac{1}{\eta_k}\big(\boldsymbol{y}^{(k)} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^{(k)} - \eta_k\boldsymbol{g}_k)\big)$
    until $\tilde{\boldsymbol{g}}_{\eta_k}^\top\boldsymbol{A}^\top\boldsymbol{A}\tilde{\boldsymbol{g}}_{\eta_k} \leq \frac{1}{\eta_k}\|\tilde{\boldsymbol{g}}_{\eta_k}\|^2$
    $\boldsymbol{x}^{(k+1)} = \boldsymbol{y}^{(k)} - \eta_k\tilde{\boldsymbol{g}}_{\eta_k}$
    $\theta_{k+1} = \frac{2\theta_k}{\theta_k + \sqrt{\theta_k^2+4}}$
    $\beta_{k+1} = \theta_k(1 - \theta_k)/(\theta_k^2 + \theta_{k+1})$
    $\boldsymbol{y}^{(k+1)} = \boldsymbol{x}^{(k+1)} + \beta_{k+1}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)})$
    $\eta_{k+1} = \eta_k/\alpha$
    if $\tilde{\boldsymbol{g}}_{\eta_k}^\top(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) > 0$ **then**
        $\theta_{k+1} = 0$

computational complexity per iteration as PGD and Bt-PGD[7] while achieving significantly faster convergence rate. Further details on ARNAPGD are provided in Algorithm 5.3. In the next section, we compare the performance of PGD with a fixed optimal step size, Bt-PGD, and ARNAPGD for UMLS.

[7]The number of matrix-vector products in ARNAPGD is exactly the same as that in Bt-PGD.

5.6 Numerical Evaluation

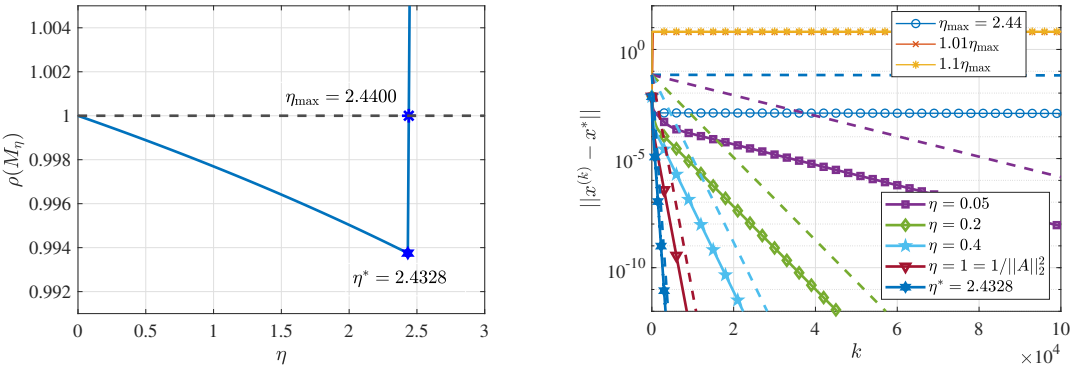
This section demonstrates the correctness of our theoretical result on the linear convergence of PGD for UMLS in Theorem 5.1. We show through numerical simulation that our predicted rate of convergence matches the decrease in the distance to the solution through iterations. Moreover, we illustrate the effectiveness of the two variants of PGD with adaptive step sizes proposed in Section 5.5. Finally, we present a simple 2-D example of the region of convergence to demonstrate our theoretical bound in (5.25).

5.6.1 PGD with a Fixed Step Size

Data generation. In the following, we create an UMLS setting in which $\boldsymbol{x}^* \in \mathcal{C}$ satisfies

$$\begin{cases} \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) = (\operatorname{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2)\boldsymbol{x}^*, \\ \boldsymbol{H} = \boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{Z} - \operatorname{diag}(\boldsymbol{\gamma}) \succ \boldsymbol{0}_N \end{cases}$$

as follows. First, we generate two matrices $\Re$ and $\Im$ of size $M \times N$, where $M = 50$ and $N = 40$, with i.i.d normally distributed ($\mathcal{N}(0,1)$) entries. The matrix $\boldsymbol{A}$ is computed from $\Re$ and $\Im$ using (5.3) Second, we generate a random vector $\boldsymbol{v} \in \mathbb{R}^N$ with $i.i.d$ normally distributed ($\mathcal{N}(0, 0.1^2)$) entries and a random vector $\boldsymbol{t} \in \{-1, 1\}^N$ with uniformly distributed entries. Then, we obtain $\boldsymbol{x}^*$ and $\boldsymbol{\gamma}$ by



(a) (b)

Figure 5.2: Convergence of PGD with a fixed step size for UMLS. (a) Plot of the convergence rate $\rho(\boldsymbol{M}_\eta)$ as a function of the step size η. The black dashed line is the line $\eta = 1$, emphasizing that the local convergence is guaranteed when $\rho(\boldsymbol{M}_\eta) < 1$. The blue start represents the maximum step size $\eta_{\max}$ such that $\rho(\boldsymbol{M}_{\eta_{\max}}) = 1$, while the blue hexagram represents the optimal step size is $\eta^* = \operatorname{argmin}_{\eta>0} \rho(\boldsymbol{M}_\eta)$. (b) Plot of the distance between the current update and the local minimum as a function of the number of iterations for various fixed step sizes. Dashed lines represent the corresponding upper bounds with exponential decay, i.e., $\rho^k(\boldsymbol{M}_\eta)$ up to a constant.

setting

$$\begin{cases} \gamma_i = t_i \left\| \boldsymbol{S}_i(\boldsymbol{A}^\top \boldsymbol{v}) \right\| \\ \boldsymbol{S}_i(\boldsymbol{x}^*) = \boldsymbol{S}_i(\boldsymbol{A}^\top \boldsymbol{v})/\gamma_i \end{cases} \quad \text{for } i = 1, \ldots, N.$$

Next, the matrices $\boldsymbol{Z}$ and $\boldsymbol{H}$ are obtained by (5.15) and (5.16), respectively. If $\boldsymbol{H}$ is not PD, we re-run the foregoing generation process multiple times until $\boldsymbol{H} \succ \boldsymbol{0}_N$. This guarantees Condition (C1) in Theorem 5.1 is satisfied. Finally, we compute $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{v}$ and initialize $\boldsymbol{x}^{(0)}$ near $\boldsymbol{x}^*$ by adding a random vector with $i.i.d$ normally distributed ($\mathcal{N}(0, 0.001^2)$) entries.

Results. Figure 5.2(a) demonstrates the convergence rate $\rho(\boldsymbol{M}_\eta)$ (blue solid line) as a function of the step size η. Recall that $\boldsymbol{M}_\eta = \boldsymbol{I}_N - \eta(\boldsymbol{I}_N - \eta \operatorname{diag}(\boldsymbol{\gamma}))^{-1}\boldsymbol{H}$ and hence, $\rho(\boldsymbol{M}_\eta)$ is a non-linear function of η. It can be seen from the plot that $\rho(\boldsymbol{M}_\eta)$ approaches 1 (slow convergence) when η approaches either 0 or $\eta_{\max} = 2.44$. The optimal step size that yields the fastest convergence for PGD with a fixed step size is $\eta^* = \operatorname{argmin}_{\eta>0} \rho(\boldsymbol{M}_\eta) = 2.4328$. Figure 5.2(b) shows the convergence of PGD with various fixed step sizes. We observe that for $\eta > \eta_{\max}$ (the red and yellow solid lines), the algorithm diverges from the designed strict local minimum $\boldsymbol{x}^*$. For step sizes less than $\eta_{\max}$, our theoretical rate (dashed lines) matches well with the empirical rate (solid lines). Moreover, PGD with the optimal step size η^* converges roughly twice as fast as one with the step size $\eta = 1/\left\|\boldsymbol{A}\right\|_2^2$ proposed in [206], suggesting that the latter choice, while being commonly used in the literature, is conservative.

5.6.2 Adaptive Schemes for Step Size

To illustrate the role of α in exploring larger step sizes with faster convergence while balancing the cost of backtracking steps, we plot the error through iterations $\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|$ against the number of matrix-vector products, which dominates the computational complexity per iteration, in Fig. 5.3. The data used in this simulation is the same as in the previous section. While the smaller values of α seems to yields faster convergence (see Fig. 5.3(a)), they indeed require more backtracking steps at each iteration (see Fig. 5.3(b)). As a result, the overall computation is higher for smaller values of α. It can be seen from Fig. 5.3(c) that the best choice of α is $\alpha = \beta = 0.8$. In addition, we observe that the total cost of Bt-PGD is comparable to that of PGD with the optimal fixed step size. However, Bt-PGD does not use any prior knowledge about the solution $\boldsymbol{x}^*$. Finally, Fig. 5.3(d) shows the fluctuation in the step size η around the optimal step size $\eta^* = 2.4328$. It is interesting to note that even though $\eta > \eta_{\max}$ at some iterations, the algorithm is able to converge to the designed local minimum $\boldsymbol{x}^*$.

Figure 5.4 depicts the fast convergence of ARNAPGD compared to PGD and Bt-PGD. The data used in this simulation is the same as in the previous section. Finally, we note that both of the foregoing adaptive schemes do not come with convergence guarantees in our setting since $\mathcal{C}$ is non-convex. Nonetheless, they do not require prior knowledge of the solution and their effectiveness is depicted clearly through our numerical results.

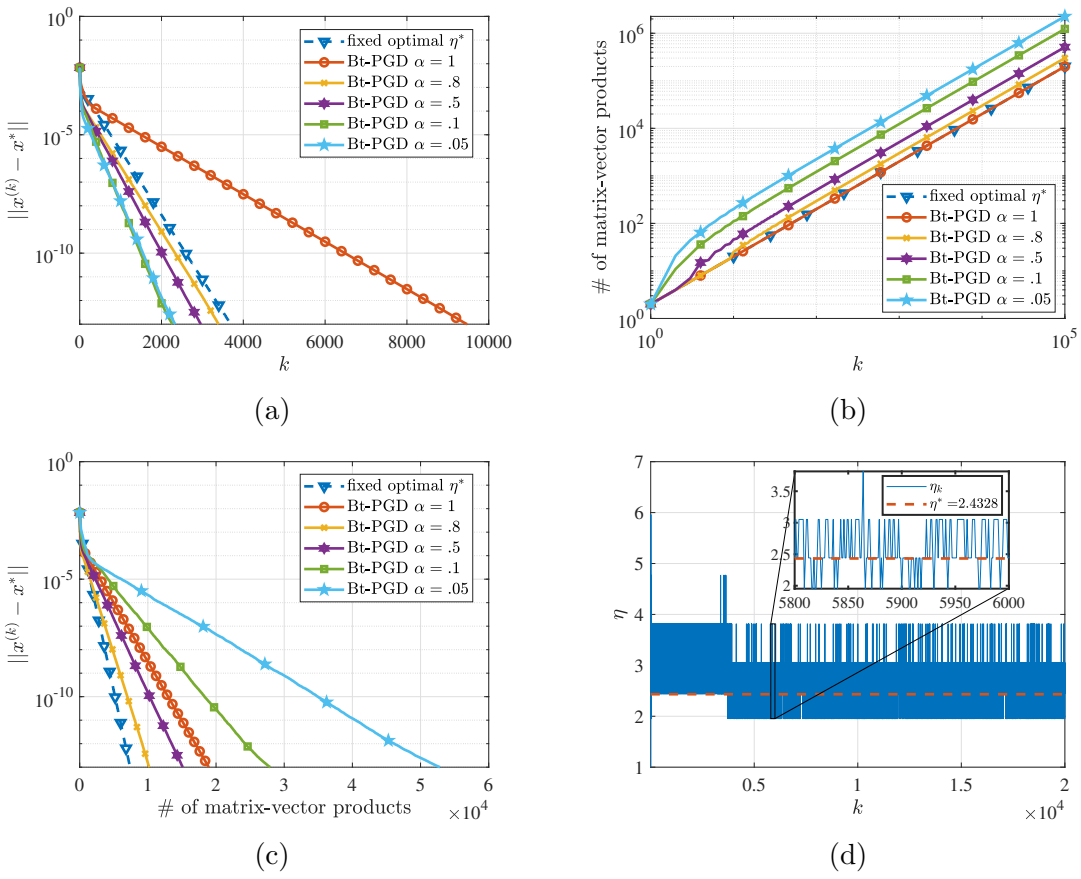


Figure 5.3: Convergence of Bt-PGD with various values of α and a fixing value of $\beta = 0.8$. (a) Plot of the distance from the current update of Bt-PGD to the local minimum as a function of the number of iterations. A dashed blue line is included as an illustration of the convergence of PGD with the fixed optimal step size η^*. (b) Plot of the number of matrix-vector products used by Bt-PGD as a function of the number of iterations. (c) Plot of the distance from the current update of Bt-PGD to the local minimum as a function of the number of matrix-vector products. (d) Plot of the change in the backtracking step size η through the first 20000 iterations for Bt-PGD with $\alpha = \beta = 0.8$. A zoomed plot is included on top of the original plot for enhanced visualization. After a few thousand iterations, we observe that the adaptive step size η_k fluctuates around the optimal step size $\eta^* = 2.4328$ (red dashed line).

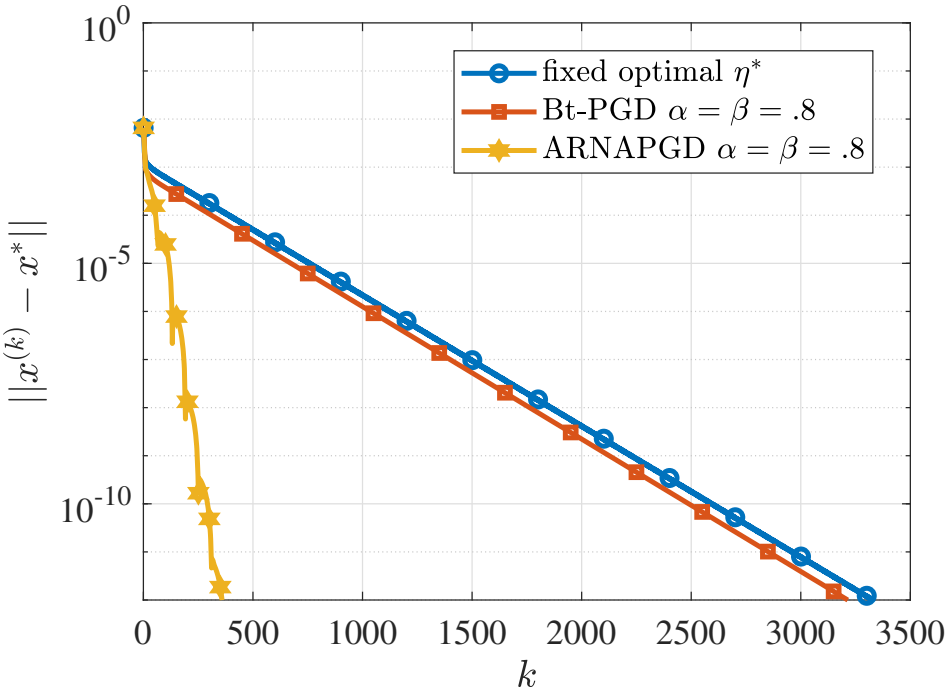


Figure 5.4: Plot of the distance from the current updates of PGD with the fixed optimal step size η^*, Bt-PGD with $\alpha = \beta = 0.8$, and ARNAPGD to the local minimum $\boldsymbol{x}^*$ as a function of the number of iterations. It is highlighted that ARNAPGD outperforms the other two algorithms significantly while remaining similar computational complexity per iteration.

5.6.3 Region of Convergence

In this subsection, we demonstrate the region of local convergence for PGD in a 2-D setting. Since $N = 1$ in this case, the constraint set $\mathcal{C}$ is indeed a 2-D circle. As can be seen from Fig. 5.5, the least-squares objective has an unconstrained global minimum at $\boldsymbol{x}^*_{unc} = [0.7, 0.2]^\top$, with $\boldsymbol{A} = \text{diag}([5, 1])$ and $\boldsymbol{b} = [3.5, 0.2]^\top$. Using Lemma 5.1, we can find the four stationary points of the 2-D UMLS problem by solving the following system of non-linear equations

$$\begin{cases} x_1^2 + x_2^2 = 1 \\ 25x_1 - 17.5 = \gamma x_1 \\ x_2 - 0.2 = \gamma x_2. \end{cases}$$

Moreover, based on the positivity of the reduced Riemannian Hessian $h = 25x_2^2 + x_1^2 - \gamma$ (which is a scalar in the 2-D setting), one can apply Lemma 5.2 to determine the two local maxima (purple hexagrams) and two local minima (green asterisk and red diamond). Additionally, for each local minima, the rate of convergence is given by $\rho_\eta = 1 - \eta h/(1 - \eta\gamma)$, with the maximum possible step size $\eta_{\max} = 2/(h + 2\gamma)$. In Fig. 5.5, we pick $\eta = 0.0755$ and compute the theoretical region of convergence for each local minima using (5.25). On the other hand, the empirical region of convergence is obtained follows. First, we run PGD with $\eta = 0.0755$ and 1000 different initialization uniformly distributed on the unit circle. Second, we check whether the algorithm stops inside the theoretical region of convergence after 1000 iterations to determine if it converges to the corresponding local minimum. Finally,

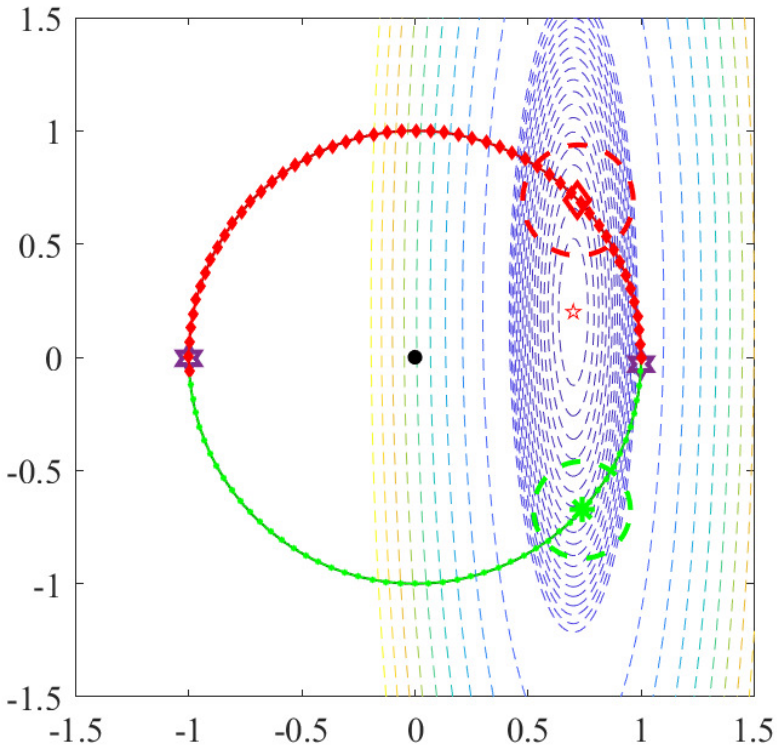


Figure 5.5: An 2-D illustration of the region of convergence given by the constant $c_0(\boldsymbol{x}^*, \eta)$ in (5.25). On the circle, the two purple hexagrams denote the local maxima, while the green asterisk and the red diamond denote the local minima of the problem. The red star located inside the circle is the solution to the unconstrained least squares. For a given fixed step size η, each local minimum is associated with (i) an estimated region of convergence (dashed circle) given by $c_0(\boldsymbol{x}^*, \eta)$ and (ii) an empirical region of convergence (circular arc with matching color) given by running PGD with the fixed step size η and initialization at a given point on the circle to verify which local minimum it converges to.

we color the initialization points by the color of the corresponding local minimum PGD converges to (either green or red). While Fig. 5.5 verifies that our theoretical region of convergence falls inside the empirical region of convergence, it also reveals that our bound is conservative in this example.

5.7 Conclusion and Future Work

We performed a novel analysis of linear convergence of projected gradient descent for the unit-modulus least-squares problem. Our analysis reveals that near the solution, the convergence is actually linear instead of sublinear. Moreover, we identified the sufficient conditions for linear convergence and provided an exact expression of the linear convergence rate. The theoretical rate predicts accurately the asymptotic convergence of PGD for UMLS in our numerical simulation. On the practical side, we propose two variant of PGD with adaptive step sizes that obtain fast convergence without prior knowledge about the solution.

For future work, we plan to improve our bound on the region of convergence. This requires further investigation into the bounding techniques used in the proof of Theorem 5.1. Another potential direction is to develop the analysis for linear convergence of Bt-PGD and ARNAPGD. While convergence guarantees for backtracking line search and Nesterov's accelerated gradient have been proposed in the optimization literature [24,160], they often involve the spectral radius that depends linearly on the step size η. The UMLS problem, on the other hand, involve the spectral radius $\rho(\boldsymbol{M}_\eta)$ that depends non-linearly on η. This makes it challenging

for determining closed-form expressions of the optimal step size in both plain PGD and accelerated PGD.

5.8 Appendix

5.8.1 Proof of Lemma 5.2

Since the constraint gradients are of form $\{\boldsymbol{e}_i \otimes \boldsymbol{S}_i(\boldsymbol{x}^*)\}_{i=1}^N$, the tangent space to $\mathcal{C}$ at $\boldsymbol{x}^*$ is given by

$$T_{\boldsymbol{x}^*}\mathcal{C} = \Big\{\boldsymbol{y} \in \mathbb{R}^{2N} \mid \big(\sum_{i=1}^{N} \boldsymbol{e}_i\boldsymbol{e}_i^\top \otimes \boldsymbol{S}_i(\boldsymbol{x}^*)\big)^\top \boldsymbol{y} = \boldsymbol{0}_N\Big\}.$$

Denote $\boldsymbol{v}_i = [-x_{2i}^*, x_{2i-1}^*]^\top$ for $i = 1, \ldots, N$. A basis of $T_{\boldsymbol{x}^*}\mathcal{C}$ is given by $\{\boldsymbol{e}_i \otimes \boldsymbol{v}_i\}_{i=1}^N$, i.e., the columns of $\boldsymbol{Z}$. Alternatively, $T_{\boldsymbol{x}^*}\mathcal{C}$ can be represented as

$$T_{\boldsymbol{x}^*}\mathcal{C} = \big\{\boldsymbol{Z}\boldsymbol{z} \mid \boldsymbol{z} \in \mathbb{R}^N\big\}. \tag{5.29}$$

($\Rightarrow$) From Chapter 11.5 in [140], the second-order necessary condition for a stationary point $\boldsymbol{x}^*$ to be a local minimum of (5.6) is $\boldsymbol{y}^\top \nabla_{\boldsymbol{x}}^2 L(\boldsymbol{x}^*, \boldsymbol{\gamma})\boldsymbol{y} \geq 0$ for all

$\boldsymbol{y} \in T_{\boldsymbol{x}^*}\mathcal{C}$. In other words, for any $\boldsymbol{z} \in \mathbb{R}^N$, we have

$$\begin{aligned} 0 &\leq (\boldsymbol{Z}\boldsymbol{z})^\top \big(\boldsymbol{A}^\top \boldsymbol{A} - \operatorname{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2\big)(\boldsymbol{Z}\boldsymbol{z}) \\ &= \boldsymbol{z}^\top \big(\boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{Z} - \boldsymbol{Z}^\top (\operatorname{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2)\boldsymbol{Z}\big)\boldsymbol{z} \\ &= \boldsymbol{z}^\top \big(\boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{Z} - \boldsymbol{Z}^\top \boldsymbol{Z} \operatorname{diag}(\boldsymbol{\gamma})\big)\boldsymbol{z} \\ &= \boldsymbol{z}^\top \big(\boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{Z} - \operatorname{diag}(\boldsymbol{\gamma})\big)\boldsymbol{z}, \end{aligned}$$

where the second equality stems from Lemma 5.11 and the third equality uses the semi-orthogonality of $\boldsymbol{Z}$. Thus, we conclude that $\boldsymbol{H} \succeq \boldsymbol{0}_N$.

($\Leftarrow$) From Chapter 11.5 in [140], the second-order sufficient condition for a stationary point $\boldsymbol{x}^*$ to be a local minimum of (5.6) is $\boldsymbol{y}^\top \nabla^2_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\gamma})\boldsymbol{y} > 0$ for all $\boldsymbol{y} \in T_{\boldsymbol{x}^*}\mathcal{C}$. By the same argument, this is equivalent to $\boldsymbol{H} \succ \boldsymbol{0}_N$.

5.8.2 Proof of Remark 5.1

Recall that the objective function is given by $f = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2/2$. By definition of the Riemannian Hessian [124], for any vector fields $U, V : \mathcal{C} \to T\mathcal{C}$ on $\mathcal{C}$, we have

$$\operatorname{Hess} f(U, V) = \langle \nabla_U \operatorname{grad} f, V \rangle, \tag{5.30}$$

where $\operatorname{grad} f : \mathcal{C} \to T\mathcal{C}$ is the Riemannian gradient given by

$$\operatorname{grad} f(\boldsymbol{x}) = \boldsymbol{Z}\boldsymbol{Z}^\top \nabla f(\boldsymbol{x}) = \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}), \tag{5.31}$$

for $\boldsymbol{x} \in \mathcal{C}$ and $\boldsymbol{Z}$ is the corresponding basis matrix of the tangent space to $\mathcal{C}$ at $\boldsymbol{x}$ (see Lemma 5.2). In addition, $\nabla_U \mathrm{grad} f$ is the covariant derivative of the vector field $\mathrm{grad} f$ in the direction of the vector field U. It is fact that the covariant derivative is the orthogonal projection of the directional derivative onto the tangent space of the manifold, i.e.,

$$\begin{aligned} \nabla_U \mathrm{grad} f(\boldsymbol{x}) &= \boldsymbol{Z}\boldsymbol{Z}^\top D_U \mathrm{grad} f(\boldsymbol{x}) \\ &= \boldsymbol{Z}\boldsymbol{Z}^\top \lim_{t \to 0} \frac{\mathrm{grad} f(\boldsymbol{x} + t\boldsymbol{u}) - \mathrm{grad} f(\boldsymbol{x})}{t}, \end{aligned} \tag{5.32}$$

where $\boldsymbol{u} = U(\boldsymbol{x})$. Substituting (5.31) into the numerator on the RHS of (5.32) and simplifying the expression, we obtain

$$\nabla_U \mathrm{grad} f(\boldsymbol{x}) = \boldsymbol{Z}\boldsymbol{Z}^\top \big(\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{u} - \boldsymbol{B}\boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \big),$$

where

$$\boldsymbol{B} = \sum_{i=1}^{N} \boldsymbol{e}_i \boldsymbol{e}_i^\top \otimes \Big(\boldsymbol{S}_i(\boldsymbol{u}) \big(\boldsymbol{S}_i(\boldsymbol{x}) \big)^\top + \boldsymbol{S}_i(\boldsymbol{x}) \big(\boldsymbol{S}_i(\boldsymbol{u}) \big)^\top \Big).$$

Now, denoting $\boldsymbol{v} = V(\boldsymbol{x})$ and evaluating (5.30) at $\boldsymbol{x}$ yields

$$\begin{aligned} \mathrm{Hess} f_{\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{v}) &= \boldsymbol{v}^\top \boldsymbol{Z}\boldsymbol{Z}^\top \big(\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{u} - \boldsymbol{B}\boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \big) \\ &= \boldsymbol{v}^\top \big(\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{u} - \boldsymbol{B}\boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \big), \end{aligned} \tag{5.33}$$

where the last equality stems from $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{C}$ and hence, $\boldsymbol{v} = \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{v}$. In the case $\boldsymbol{x} = \boldsymbol{x}^*$ is a stationary point of (5.6) with the Lagrange multiplier $\boldsymbol{\gamma}$, one can substituting (5.14) into (5.33) to obtain

$$\mathrm{Hess} f_{\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{v}^\top \big(\boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{u} - \boldsymbol{B}(\mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2)\boldsymbol{x} \big). \tag{5.34}$$

Notice that $\boldsymbol{x} = \sum_{i=1}^N \boldsymbol{e}_i \otimes \boldsymbol{S}_i(\boldsymbol{x})$ and $\big(\boldsymbol{S}_i(\boldsymbol{u})\big)^\top \boldsymbol{S}_i(\boldsymbol{x}) = 0$ for all $i = 1, \ldots, N$. Therefore, the second term on the RHS of (5.34) can be simplified as

$$\begin{aligned} \boldsymbol{B}(\mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2)\boldsymbol{x} &= \sum_{i=1}^N \gamma_i \boldsymbol{e}_i \otimes \boldsymbol{S}_i(\boldsymbol{u}) \\ &= (\mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2)\boldsymbol{u}. \end{aligned}$$

Substituting back into (5.34) and reorganizing terms, we obtain the Riemannian Hessian as

$$\mathrm{Hess} f_{\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \big(\boldsymbol{A}^\top \boldsymbol{A} - (\mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2) \big) \boldsymbol{v}. \tag{5.35}$$

Finally, it follows from (5.29) that there is an one-to-one correspondence between the tangent space $T_{\boldsymbol{x}}\mathcal{C}$ and $\mathbb{R}^N$, i.e., $\boldsymbol{u} = \boldsymbol{Z}\tilde{\boldsymbol{u}}$ and $\boldsymbol{v} = \boldsymbol{Z}\tilde{\boldsymbol{v}}$ for $\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}} \in \mathbb{R}^N$. Hence,

we can define a bilinear function $H : \mathbb{R}^N \otimes \mathbb{R}^N \to \mathbb{R}$:

$$\begin{aligned} H(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) &\triangleq \mathrm{Hess} f_{\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{v}) \\ &= (\boldsymbol{Z}\tilde{\boldsymbol{u}})^\top \big(\boldsymbol{A}^\top \boldsymbol{A} - (\mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2)\big)(\boldsymbol{Z}\tilde{\boldsymbol{v}}) \\ &= \tilde{\boldsymbol{u}}^\top (\boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{Z} - \mathrm{diag}(\boldsymbol{\gamma}))\tilde{\boldsymbol{v}}, \end{aligned}$$

where the last equality stems from $\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{I}_N$. In other words, $\mathrm{Hess} f_{\boldsymbol{x}}$ admits a compact matrix representation

$$\boldsymbol{H} = \boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{Z} - \mathrm{diag}(\boldsymbol{\gamma}).$$

5.8.3 Proof of Lemma 5.3

($\Rightarrow$) Assume $\boldsymbol{x}^*$ is a fixed point of Algorithm 5.1 with step size $\eta > 0$, i.e.,

$$\boldsymbol{x}^* = \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* - \eta \boldsymbol{r}), \tag{5.36}$$

where $\boldsymbol{r} = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})$. We will show there exists $\boldsymbol{\gamma} \in \mathbb{R}^N$ such that for all $i = 1, \ldots, N$,

$$\boldsymbol{S}_i(\boldsymbol{r}) = \gamma_i \boldsymbol{S}_i(\boldsymbol{x}^*) \tag{5.37}$$

and

$$\begin{cases} \gamma_i < 1/\eta & \text{if } \boldsymbol{S}_i(\boldsymbol{x}^*) \neq \boldsymbol{s}, \\ \gamma_i \leq 1/\eta & \text{if } \boldsymbol{S}_i(\boldsymbol{x}^*) = \boldsymbol{s}, \end{cases} \tag{5.38}$$

where we recall that $\boldsymbol{s} = [1, 0]^\top$.

For $i = 1, \ldots, N$, applying the 2-selection operator $\boldsymbol{S}_i(\cdot)$ to both side of (5.36) and substituting the RHS by the definition of $\mathcal{P}_{\mathcal{C}}$ in (5.9) yield

$$\boldsymbol{S}_i(\boldsymbol{x}^*) = \begin{cases} \frac{\boldsymbol{S}_i(\boldsymbol{x}^* - \eta \boldsymbol{r})}{\|\boldsymbol{S}_i(\boldsymbol{x}^* - \eta \boldsymbol{r})\|} & \text{if } \boldsymbol{S}_i(\boldsymbol{x}^* - \eta \boldsymbol{r}) \neq \boldsymbol{0}_2, \\ \boldsymbol{s} & \text{if } \boldsymbol{S}_i(\boldsymbol{x}^* - \eta \boldsymbol{r}) = \boldsymbol{0}_2. \end{cases} \tag{5.39}$$

We split (5.39) into two cases based on the value of $\boldsymbol{S}_i(\boldsymbol{x}^*)$. If $\boldsymbol{S}_i(\boldsymbol{x}^*) \neq \boldsymbol{s}$, then (5.39) implies

$$\boldsymbol{S}_i(\boldsymbol{x}^*) = \frac{\boldsymbol{S}_i(\boldsymbol{x}^* - \eta \boldsymbol{r})}{\|\boldsymbol{S}_i(\boldsymbol{x}^* - \eta \boldsymbol{r})\|} = \frac{\boldsymbol{S}_i(\boldsymbol{x}^*) - \eta \boldsymbol{S}_i(\boldsymbol{r})}{\|\boldsymbol{S}_i(\boldsymbol{x}^* - \eta \boldsymbol{r})\|},$$

which in turns can be reorganized as $\boldsymbol{S}_i(\boldsymbol{r}) = \gamma_i \boldsymbol{S}_i(\boldsymbol{x}^*)$ for

$$\gamma_i = \frac{1 - \|\boldsymbol{S}_i(\boldsymbol{x}^*) - \eta \boldsymbol{S}_i(\boldsymbol{r})\|}{\eta} < \frac{1}{\eta}. \tag{5.40}$$

If $\boldsymbol{S}_i(\boldsymbol{x}^*) = \boldsymbol{s}$, we consider two sub-cases:

1. If $\boldsymbol{S}_i(\boldsymbol{x}^* - \eta \boldsymbol{r}) \neq \boldsymbol{0}_2$, then by the same argument as the previous case, we obtain (5.40).

2. If $\boldsymbol{S}_i(\boldsymbol{x}^* - \eta\boldsymbol{r}) = \boldsymbol{0}_2$, then using the linearity of $\boldsymbol{S}_i$, we have $\boldsymbol{S}_i(\boldsymbol{r}) = \gamma_i \boldsymbol{S}_i(\boldsymbol{x}^*)$ where $\gamma_i = 1/\eta$.

In all cases, we have (5.37) and (5.38) hold. Finally, we note that the stationarity condition (5.14) is equivalent to $\boldsymbol{S}_i(\boldsymbol{r}) = \gamma_i \boldsymbol{S}_i(\boldsymbol{x}^*)$ for all $i = 1, \ldots, N$.

($\Leftarrow$) Assume $\boldsymbol{x}^*$ is a stationary point of (5.6) (i.e., (5.37 holds for all $i = 1, \ldots, N$) with the corresponding Lagrange multiplier $\boldsymbol{\gamma}$ satisfying (5.38) for all $i = 1, \ldots, N$. We will prove (5.36) by showing that

$$\boldsymbol{S}_i\big(\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* - \eta\boldsymbol{r})\big) = \boldsymbol{S}_i(\boldsymbol{x}^*), \tag{5.41}$$

for any $i = 1, \ldots, N$.

By the definition of $\mathcal{P}_{\mathcal{C}}$ in (5.9), we have

$$\boldsymbol{S}_i\big(\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* - \eta\boldsymbol{r})\big) = \begin{cases} \frac{\boldsymbol{S}_i(\boldsymbol{x}^* - \eta\boldsymbol{r})}{\|\boldsymbol{S}_i(\boldsymbol{x}^* - \eta\boldsymbol{r})\|} & \text{if } \boldsymbol{S}_i(\boldsymbol{x}^* - \eta\boldsymbol{r}) \neq \boldsymbol{0}_2, \\ \boldsymbol{s} & \text{if } \boldsymbol{S}_i(\boldsymbol{x}^* - \eta\boldsymbol{r}) = \boldsymbol{0}_2. \end{cases} \tag{5.42}$$

Using the linearity of $\boldsymbol{S}_i(\cdot)$ and then the stationarity condition in (5.37) yield

$$\begin{aligned} \boldsymbol{S}_i(\boldsymbol{x}^* - \eta\boldsymbol{r}) &= \boldsymbol{S}_i(\boldsymbol{x}^*) - \eta\boldsymbol{S}_i(\boldsymbol{r}) \\ &= \boldsymbol{S}_i(\boldsymbol{x}^*) - \eta\gamma_i\boldsymbol{S}_i(\boldsymbol{x}^*) = (1 - \eta\gamma_i)\boldsymbol{S}_i(\boldsymbol{x}^*). \end{aligned} \tag{5.43}$$

Since $\boldsymbol{x} \in \mathcal{C}$, $\|\boldsymbol{S}_i(\boldsymbol{x}^*)\| = 1$. Taking the norm of both sides in (5.43) and using

(5.38) to remove the absolute value, we obtain

$$\begin{aligned}\|\boldsymbol{S}_i(\boldsymbol{x}^* - \eta\boldsymbol{r})\| &= \|(1-\eta\gamma_i)\boldsymbol{S}_i(\boldsymbol{x}^*)\| \\ &= |1-\eta\gamma|\,\|\boldsymbol{S}_i(\boldsymbol{x}^*)\| = 1-\eta\gamma.\end{aligned}$$

Therefore, (5.42) is equivalent to

$$\boldsymbol{S}_i\big(\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* - \eta\boldsymbol{r})\big) = \begin{cases} \boldsymbol{S}_i(\boldsymbol{x}^*) & \text{if } 1-\eta\gamma_i \neq 0, \\ \boldsymbol{s} & \text{if } 1-\eta\gamma_i = 0. \end{cases} \tag{5.44}$$

- If $1-\eta\gamma_i \neq 0$, then (5.41) holds trivially.
- If $1-\eta\gamma_i = 0$, then $\boldsymbol{S}_i(\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* - \eta\boldsymbol{r})) = \boldsymbol{s}$ and $\gamma_i = 1/\eta$. From (5.38), the latter only holds if $\boldsymbol{S}_i(\boldsymbol{x}^*) = \boldsymbol{s}$. Thus, we obtain $\boldsymbol{S}_i(\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* - \eta\boldsymbol{r})) = \boldsymbol{S}_i(\boldsymbol{x}^*) = \boldsymbol{s}$.

In both case, we have (5.41) holds for all $i = 1, \ldots, N$. This completes our proof of the lemma.

5.8.4 Proof of Proposition 5.1

The proof of this lemma is based on the following result for the projection onto the unit sphere [217]:

Lemma 5.10. *(Rephrased from Lemma 5 in [217]) Let $\boldsymbol{x}$ be a point on the unit*

sphere $\mathcal{S}^{n-1}$. Then, for any $\boldsymbol{\delta} \in \mathbb{R}^n$, the projection onto $\mathcal{S}^{n-1}$ satisfies

$$\mathcal{P}_{\mathcal{S}^{n-1}}(\boldsymbol{x}+\boldsymbol{\delta}) = \boldsymbol{x} + \big(\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^\top\big)\boldsymbol{\delta} + \boldsymbol{q}_{\mathcal{S}^{n-1}}(\boldsymbol{\delta}), \tag{5.45}$$

where $\|\boldsymbol{q}_{\mathcal{S}^{n-1}}(\boldsymbol{\delta})\| \leq 2\,\|\boldsymbol{\delta}\|^2$.

Applying Lemma 5.10 to the unit circle $\mathcal{S}^1$ (corresponding to the case $n = 2$), we have, for each $i = 1, \ldots, N$,

$$\begin{aligned}
\boldsymbol{S}_i\big(\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})\big) &= \mathcal{P}_{\mathcal{S}^1}\big(\boldsymbol{S}_i(\boldsymbol{x}+\boldsymbol{\delta})\big) \\
&= \mathcal{P}_{\mathcal{S}^1}\big(\boldsymbol{S}_i(\boldsymbol{x}) + \boldsymbol{S}_i(\boldsymbol{\delta})\big) \\
&= \boldsymbol{S}_i(\boldsymbol{x}) + \big(\boldsymbol{I}_2 - \boldsymbol{S}_i(\boldsymbol{x})(\boldsymbol{S}_i(\boldsymbol{x}))^\top\big)\boldsymbol{S}_i(\boldsymbol{\delta}) + \boldsymbol{q}_{\mathcal{S}^1}\big(\boldsymbol{S}_i(\boldsymbol{\delta})\big) \\
&= \boldsymbol{S}_i(\boldsymbol{x}) + \boldsymbol{v}_i\boldsymbol{v}_i^\top\boldsymbol{S}_i(\boldsymbol{\delta}) + \boldsymbol{q}_{\mathcal{S}^1}\big(\boldsymbol{S}_i(\boldsymbol{\delta})\big),
\end{aligned}$$

where $\boldsymbol{v}_i = [-x_{2i}, x_{2i-1}]^\top$. Using the property of the 2-selection operator in (5.4), we further have

$$\begin{aligned}
\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta}) &= \sum_{i=1}^{N} \boldsymbol{e}_i \otimes \boldsymbol{S}_i\big(\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta})\big) \\
&= \sum_{i=1}^{N} \boldsymbol{e}_i \otimes \Big(\boldsymbol{S}_i(\boldsymbol{x}) + \boldsymbol{v}_i\boldsymbol{v}_i^\top\boldsymbol{S}_i(\boldsymbol{\delta}) + \boldsymbol{q}_{\mathcal{S}^1}\big(\boldsymbol{S}_i(\boldsymbol{\delta})\big)\Big) \\
&= \sum_{i=1}^{N} \boldsymbol{e}_i \otimes \boldsymbol{S}_i(\boldsymbol{x}) + \sum_{i=1}^{N} (\boldsymbol{e}_i \otimes \boldsymbol{v}_i\boldsymbol{v}_i^\top)\boldsymbol{S}_i(\boldsymbol{\delta}) + \sum_{i=1}^{N} \boldsymbol{e}_i \otimes \boldsymbol{q}_{\mathcal{S}^1}\big(\boldsymbol{S}_i(\boldsymbol{\delta})\big) \\
&= \boldsymbol{x} + \boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{\delta} + \boldsymbol{q}(\boldsymbol{\delta}),
\end{aligned} \tag{5.46}$$

where $\boldsymbol{q}(\boldsymbol{\delta})$ satisfies $\boldsymbol{S}_i(\boldsymbol{q}(\boldsymbol{\delta})) = \boldsymbol{q}_{\mathcal{S}^1}(\boldsymbol{S}_i(\boldsymbol{\delta}))$ and

$$\begin{aligned}\|\boldsymbol{q}(\boldsymbol{\delta})\|^2 &= \sum_{i=1}^{N} \|\boldsymbol{S}_i(\boldsymbol{q}(\boldsymbol{\delta}))\|^2 = \sum_{i=1}^{N} \|\boldsymbol{q}_{\mathcal{S}^1}(\boldsymbol{S}_i(\boldsymbol{\delta}))\|^2 \\ &\leq \sum_{i=1}^{N} \big(2\,\|\boldsymbol{S}_i(\boldsymbol{\delta})\|^2\big)^2 \leq \big(\sum_{i=1}^{N} 2\,\|\boldsymbol{S}_i(\boldsymbol{\delta})\|^2\big)^2 \\ &= 4\big(\sum_{i=1}^{N}(\delta_{2i-1}^2 + \delta_{2i}^2)\big)^2 = 4\big(\sum_{j=1}^{2N} \delta_j^2\big)^2 = 4\,\|\boldsymbol{\delta}\|^4 .\end{aligned}$$

This completes our proof of the lemma.

5.8.5 Proof of Lemma 5.4

In this section, we show that when Conditions (C1) and (C2) in Theorem 5.1 hold, Condition (C3'), i.e.,

$$\eta(\lambda_1(\boldsymbol{H}) + 2\gamma_i) < 2, \tag{5.47}$$

for all $i = 1, \ldots, N$, is sufficient for Condition (C3). First, we prove that $\boldsymbol{D}_\eta = (\boldsymbol{I}_N - \eta\,\mathrm{diag}(\boldsymbol{\gamma}))^{-1}$ is PSD. Second, we show that all the eigenvalues of $\boldsymbol{D}_\eta\boldsymbol{H}$ lie between 0 and $(1-\eta\gamma_i)^{-1}\lambda_1(\boldsymbol{H})$ (exclusively). Third, we claim that the spectral radius of $\boldsymbol{M}_\eta = \boldsymbol{I}_N - \eta\boldsymbol{D}_\eta\boldsymbol{H}$ is strictly less than 1.

In the first step, rearranging (5.47), we obtain $\eta\lambda_1(\boldsymbol{H})/2 < 1 - \eta\gamma_i$. By Condition (C2), we have $\lambda_1(\boldsymbol{H}) > 0$. Since $\eta > 0$, it follows that $0 < \eta\lambda_1(\boldsymbol{H})/2 < 1-\eta\gamma_i$. Thus, the diagonal matrix $\boldsymbol{D}_\eta$ has all positive entries and hence, is a PSD matrix.

In the second step, we use the inequalities for the eigenvalues of the product of two PSD matrices in [223] to obtain

$$\lambda_i(\boldsymbol{D}_\eta)\lambda_N(\boldsymbol{H}) \leq \lambda_i(\boldsymbol{D}_\eta\boldsymbol{H}) \leq \lambda_i(\boldsymbol{D}_\eta)\lambda_1(\boldsymbol{H}), \tag{5.48}$$

for all $i = 1, \ldots, N$. Since both $\boldsymbol{D}_\eta$ and $\boldsymbol{H}$ are PSD, we can lower bound the eigenvalues of $\boldsymbol{D}_\eta\boldsymbol{H}$ by $\lambda_i(\boldsymbol{D}_\eta\boldsymbol{H}) \geq \lambda_i(\boldsymbol{D}_\eta)\lambda_N(\boldsymbol{H}) > 0$. On the other hand, substituting $\lambda_i(\boldsymbol{D}_\eta) = (1-\eta\gamma_i)^{-1}$ into the upper bound in (5.48) yields $\lambda_i(\boldsymbol{D}_\eta\boldsymbol{H}) \leq (1-\eta\gamma_i)^{-1}\lambda_1(\boldsymbol{H})$. Finally, using the fact that $\lambda_i(\boldsymbol{M}_\eta) = 1 - \eta\lambda_i(\boldsymbol{D}_\eta\boldsymbol{H})$ and $0 < \lambda_i(\boldsymbol{D}_\eta\boldsymbol{H}) \leq (1-\eta\gamma_i)^{-1}\lambda_1(\boldsymbol{H})$, for all $i = 1, \ldots, N$, we obtain

$$1 - \frac{\eta}{1-\eta\gamma_i}\lambda_1(\boldsymbol{H}) \leq \lambda_i(\boldsymbol{M}_\eta) < 1.$$

Now, rearranging (5.47) to obtain $1-\frac{\eta}{1-\eta\gamma_i}\lambda_1(\boldsymbol{H}) > -1$, we have all the eigenvalues of $\boldsymbol{M}_\eta$ lie between -1 and 1 (exclusively). Since the spectral radius is the maximum of the absolute values of these eigenvalues, we conclude that $\rho(\boldsymbol{M}_\eta) < 1$. This completes our proof in this section.

5.8.6 Proof of Lemma 5.5

In the first part of this proof, we show that $\gamma_i < 1/\eta$ for all $i = 1, \ldots, N$. From Condition (C2), we have $\boldsymbol{D}_\eta = (\boldsymbol{I}_N - \eta\,\mathrm{diag}(\boldsymbol{\gamma}))^{-1}$ is invertible and hence, the expression of $\boldsymbol{M}_\eta$ in (5.20) is well-defined. In addition, from Condition (C1), $\boldsymbol{H}$

has a unique PD square root $\boldsymbol{H}^{1/2}$, with the inverse $\boldsymbol{H}^{-1/2}$. Thus, we have

$$\begin{aligned}\boldsymbol{H}^{1/2}\boldsymbol{M}_\eta\boldsymbol{H}^{-1/2} &= \boldsymbol{H}^{1/2}\Big(\boldsymbol{I}_N - \eta\Big(\boldsymbol{I}_N - \eta\,\mathrm{diag}(\boldsymbol{\gamma})\Big)^{-1}\boldsymbol{H}\Big)\boldsymbol{H}^{-1/2} \\ &= \boldsymbol{I}_N - \eta\boldsymbol{H}^{1/2}\boldsymbol{D}_\eta\boldsymbol{H}^{1/2} \triangleq \tilde{\boldsymbol{M}}_\eta.\end{aligned}$$

This shows that $\boldsymbol{M}_\eta$ and $\tilde{\boldsymbol{M}}_\eta$ are similar matrices with the same set of eigenvalues. Combining this with Condition (C3), we obtain $\rho(\boldsymbol{M}_\eta) = \rho(\tilde{\boldsymbol{M}}_\eta) < 1$. Since $\tilde{\boldsymbol{M}}_\eta$ is symmetric, it then holds that

$$\tilde{\boldsymbol{M}}_\eta = \boldsymbol{I}_N - \eta\boldsymbol{H}^{1/2}\boldsymbol{D}_\eta\boldsymbol{H}^{-1/2} \prec \boldsymbol{I}_N,$$

which in turn yields $\boldsymbol{H}^{1/2}\boldsymbol{D}_\eta\boldsymbol{H}^{1/2} \succ \boldsymbol{0}_N$. By the definition of PD matrices, for any vector $\boldsymbol{u} \in \mathbb{R}^N$, it holds that $\boldsymbol{u}^\top\boldsymbol{H}^{1/2}\boldsymbol{D}_\eta\boldsymbol{H}^{1/2}\boldsymbol{u} > 0$. Alternatively, we can write $\boldsymbol{v}^\top\boldsymbol{D}_\eta\boldsymbol{v} > 0$, where $\boldsymbol{v} = \boldsymbol{H}^{1/2}\boldsymbol{u}$. Notice that the mapping between $\boldsymbol{u}$ and $\boldsymbol{v}$ is bijection, which means $\boldsymbol{v}^\top\boldsymbol{D}_\eta\boldsymbol{v} > 0$ also holds for any $\boldsymbol{v} \in \mathbb{R}^N$. Consequently, $\boldsymbol{D}_\eta = \mathrm{diag}([(1-\eta\gamma_1)^{-1}, \ldots, (1-\eta\gamma_N)^{-1}])$ must be a PD matrix. Equivalently, we have $\gamma_i < 1/\eta$ for all $i = 1, \ldots, N$.

For the second part of the proof, we note that $\gamma_i < 1/\eta$, for all $i = 1, \ldots, N$, are sufficient conditions for the Lagrange multiplier condition (5.18) in Lemma 5.3. Since a strict local minimum is also a stationary point of (5.6), $\boldsymbol{x}^*$ must be a fixed point of Algorithm 5.1 with the given step size η. This completes our proof of the lemma.

5.8.7 Proof of Lemma 5.6

Using the PGD update in (5.10) and rewriting $\boldsymbol{x}^{(k)} = \boldsymbol{x}^* + \boldsymbol{\delta}^{(k)}$), we derive a recursion on the error vector as follows

$$\begin{aligned}
\boldsymbol{\delta}^{(k+1)} &= \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^* \\
&= \mathcal{P}_{\mathcal{C}}\Big(\boldsymbol{x}^{(k)} - \eta \boldsymbol{A}^\top\big(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b}\big)\Big) - \boldsymbol{x}^* \\
&= \mathcal{P}_{\mathcal{C}}\Big((\boldsymbol{x}^* + \boldsymbol{\delta}^{(k)}) - \eta \boldsymbol{A}^\top\big(\boldsymbol{A}(\boldsymbol{x}^* + \boldsymbol{\delta}^{(k)}) - \boldsymbol{b}\big)\Big) - \boldsymbol{x}^* \\
&= \mathcal{P}_{\mathcal{C}}\Big(\big(\boldsymbol{x}^* - \eta \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b})\big) + (\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{\delta}^{(k)}\Big) - \boldsymbol{x}^*. \qquad (5.49)
\end{aligned}$$

Since $\boldsymbol{x}^*$ is a stationary point of (5.6), we have $\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) = (\mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2)\boldsymbol{x}^*$. Then, the first term inside the projection $\mathcal{P}_{\mathcal{C}}$ on the RHS of (5.49) can be represented as

$$\begin{aligned}
\boldsymbol{x}^* - \eta \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) &= \big(\boldsymbol{I}_{2N} - \eta\, \mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_2\big)\boldsymbol{x}^* \\
&= \Big(\big(\boldsymbol{I}_N - \eta\, \mathrm{diag}(\boldsymbol{\gamma})\big) \otimes \boldsymbol{I}_2\Big)\boldsymbol{x}^* \\
&= (\boldsymbol{D}_\eta^{-1} \otimes \boldsymbol{I}_2)\boldsymbol{x}^* = (\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)^{-1}\boldsymbol{x}^*.
\end{aligned}$$

where we recall that $\boldsymbol{D}_\eta = (\boldsymbol{I}_N - \eta\, \mathrm{diag}(\boldsymbol{\gamma}))^{-1} \succ \boldsymbol{0}_N$ by Lemma 5.5. Thus, we rewrite (5.49) as

$$\boldsymbol{\delta}^{(k+1)} = \mathcal{P}_{\mathcal{C}}\Big((\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)^{-1}\boldsymbol{x}^* + (\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{\delta}^{(k)}\Big) - \boldsymbol{x}^*.$$

Now let $\boldsymbol{y} = \boldsymbol{x}^* + (\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}$ and using the modulus scale-invariant property of the projection $\mathcal{P}_{\mathcal{C}}((\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)^{-1}\boldsymbol{y}) = \mathcal{P}_{\mathcal{C}}(\boldsymbol{y})$, for $\boldsymbol{D}_\eta \succ \boldsymbol{0}_N$, we further obtain

$$\boldsymbol{\delta}^{(k+1)} = \mathcal{P}_{\mathcal{C}}\Big(\boldsymbol{x}^* + (\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}\Big) - \boldsymbol{x}^*. \tag{5.50}$$

Finally, applying Proposition 5.1 with the perturbation $\boldsymbol{\delta} = (\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}$ at $\boldsymbol{x} = \boldsymbol{x}^* \in \mathcal{C}$, we have

$$\begin{aligned}
&\mathcal{P}_{\mathcal{C}}\Big(\boldsymbol{x}^* + (\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}\Big) \\
&\quad = \boldsymbol{x}^* + \boldsymbol{Z}\boldsymbol{Z}^\top (\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)} + \boldsymbol{q}\big((\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\delta}^{(k)}\big).
\end{aligned}$$

Substituting this back into (5.50) yields (5.22). This completes the proof of the lemma.

5.8.8 Proof of Lemma 5.7

Since $\boldsymbol{x}^{(k)}$ lies in $\mathcal{C}$, we can represent the error vector as

$$\begin{aligned}
\boldsymbol{\delta}^{(k)} &= \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \\
&= \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^{(k)}) - \boldsymbol{x}^* \\
&= \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* + \boldsymbol{\delta}^{(k)}) - \boldsymbol{x}^*.
\end{aligned} \tag{5.51}$$

Using Proposition 5.1, we have

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^* + \boldsymbol{\delta}^{(k)}) = \boldsymbol{x}^* + \boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{\delta}^{(k)} + \boldsymbol{q}(\boldsymbol{\delta}^{(k)}).$$

Substituting this back into the RHS of (5.51) yields

$$\boldsymbol{\delta}^{(k)} = \boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{\delta}^{(k)} + \boldsymbol{q}(\boldsymbol{\delta}^{(k)}).$$

This completes our proof of the lemma.

5.8.9 Proof of Lemma 5.8

Substituting (5.23) back into the first term on the RHS of (5.22), we have

$$\begin{aligned}\boldsymbol{\delta}^{(k+1)} = {} & \boldsymbol{Z}\boldsymbol{Z}^\top(\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{\delta}^{(k)} \\ & + \boldsymbol{Z}\boldsymbol{Z}^\top(\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{q}(\boldsymbol{\delta}^{(k)}) \\ & + \boldsymbol{q}\big((\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{\delta}^{(k)}\big). \end{aligned} \tag{5.52}$$

From Lemma 5.11 and the fact that $\boldsymbol{Z}^\top\boldsymbol{Z} = \boldsymbol{I}_N$, we can represent (5.52) as

$$\begin{aligned}\boldsymbol{\delta}^{(k+1)} &= \boldsymbol{Z}\boldsymbol{D}_\eta\boldsymbol{Z}^\top(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{\delta}^{(k)} + \hat{\boldsymbol{q}}(\boldsymbol{\delta}^{(k)}) \\ &= \boldsymbol{Z}\boldsymbol{D}_\eta(\boldsymbol{I}_N - \eta\boldsymbol{Z}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{Z})\boldsymbol{Z}^\top\boldsymbol{\delta}^{(k)} + \hat{\boldsymbol{q}}(\boldsymbol{\delta}^{(k)}), \end{aligned} \tag{5.53}$$

where $\hat{\boldsymbol{q}}(\boldsymbol{\delta}) = \boldsymbol{Z}\boldsymbol{Z}^\top(\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{q}(\boldsymbol{\delta}) + \boldsymbol{q}\big((\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{\delta}\big)$.

Recall that $\boldsymbol{H} = \boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{Z} - \operatorname{diag}(\boldsymbol{\gamma})$. Thus, (5.53) is equivalent to

$$\begin{aligned}\boldsymbol{\delta}^{(k+1)} &= \boldsymbol{Z}\boldsymbol{D}_\eta(\boldsymbol{I}_N - \eta \operatorname{diag}(\boldsymbol{\gamma}) - \boldsymbol{H})\boldsymbol{Z}^\top\boldsymbol{\delta}^{(k)} + \hat{\boldsymbol{q}}(\boldsymbol{\delta}^{(k)}) \\ &= \boldsymbol{Z}(\boldsymbol{I}_N - \eta\boldsymbol{D}_\eta\boldsymbol{H})\boldsymbol{Z}^\top\boldsymbol{\delta}^{(k)} + \hat{\boldsymbol{q}}(\boldsymbol{\delta}^{(k)}).\end{aligned}$$

By the definition of $\boldsymbol{M}_\eta$ in (5.20), the last equation is the same as (5.24).

To bound the norm of $\hat{\boldsymbol{q}}(\boldsymbol{\delta}^{(k)})$, we use the triangle inequality and the product norm inequality as follows

$$\begin{aligned}\|\hat{\boldsymbol{q}}(\boldsymbol{\delta})\| &\le \left\|\boldsymbol{Z}\boldsymbol{Z}^\top(\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{q}(\boldsymbol{\delta})\right\| + \left\|\boldsymbol{q}\big((\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{\delta}\big)\right\| \\ &\le \left\|\boldsymbol{Z}\boldsymbol{Z}^\top\right\|_2 \left\|(\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\right\|_2 \|\boldsymbol{q}(\boldsymbol{\delta})\| + \left\|\boldsymbol{q}\big((\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{\delta}\big)\right\|.\end{aligned}$$

Since $\|\boldsymbol{q}(\boldsymbol{\delta})\| \le 2\|\boldsymbol{\delta}\|$ (see Proposition 5.1) and $c_\eta = \left\|(\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\right\|_2$, we further obtain

$$\begin{aligned}\|\hat{\boldsymbol{q}}(\boldsymbol{\delta})\| &\le \left\|\boldsymbol{Z}\boldsymbol{Z}^\top\right\|_2 \cdot c_\eta \cdot 2\|\boldsymbol{\delta}\|^2 + 2\left\|(\boldsymbol{D}_\eta \otimes \boldsymbol{I}_2)(\boldsymbol{I}_{2N} - \eta\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{\delta}\right\|^2 \\ &\le 2c_\eta \left\|\boldsymbol{Z}\boldsymbol{Z}^\top\right\|_2 \|\boldsymbol{\delta}\|^2 + 2c_\eta^2\|\boldsymbol{\delta}\|^2 \\ &\le 2c_\eta\|\boldsymbol{\delta}\|^2 + 2c_\eta^2\|\boldsymbol{\delta}\|^2,\end{aligned}$$

where the last inequality stems from $\left\|\boldsymbol{Z}\boldsymbol{Z}^\top\right\|_2 \le 1$ since $\boldsymbol{Z}\boldsymbol{Z}^\top$ is an orthogonal projection matrix. This completes our proof of the lemma.

5.8.10 Proof of Lemma 5.9

The proof in this section relies on Lemmas 5.8, 5.13, and 5.12. Let $\tilde{\boldsymbol{\delta}}^{(k)} = (\boldsymbol{D}_\eta^{-1/2} \otimes \boldsymbol{I}_2)\boldsymbol{\delta}^{(k)}$. Left-multiplying both sides of (5.24) with $(\boldsymbol{D}_\eta^{-1/2} \otimes \boldsymbol{I}_2)$, we have

$$\begin{aligned}\tilde{\boldsymbol{\delta}}^{(k+1)} &= (\boldsymbol{D}_\eta^{-1/2} \otimes \boldsymbol{I}_2)\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top\boldsymbol{\delta}^{(k)} + (\boldsymbol{D}_\eta^{-1/2} \otimes \boldsymbol{I}_2)\hat{\boldsymbol{q}}(\boldsymbol{\delta}^{(k)}) \\ &= (\boldsymbol{D}_\eta^{-1/2} \otimes \boldsymbol{I}_2)\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top(\boldsymbol{D}_\eta^{1/2} \otimes \boldsymbol{I}_2)\tilde{\boldsymbol{\delta}}^{(k)} + (\boldsymbol{D}_\eta^{-1/2} \otimes \boldsymbol{I}_2)\hat{\boldsymbol{q}}\big((\boldsymbol{D}_\eta^{1/2} \otimes \boldsymbol{I}_2)\tilde{\boldsymbol{\delta}}^{(k)}\big).\end{aligned} \tag{5.54}$$

Using Lemma 5.11 and substituting $\boldsymbol{M}_\eta = \boldsymbol{I}_N - \eta\boldsymbol{D}_\eta^{-1}\boldsymbol{H}$ into the RHS of (5.54) yield

$$\begin{aligned}\tilde{\boldsymbol{\delta}}^{(k+1)} &= \boldsymbol{Z}\boldsymbol{D}_\eta^{-1/2}(\boldsymbol{I}_N - \eta\boldsymbol{D}_\eta^{-1}\boldsymbol{H})\boldsymbol{D}_\eta^{1/2}\boldsymbol{Z}^\top\tilde{\boldsymbol{\delta}}^{(k)} + \tilde{\boldsymbol{q}}(\tilde{\boldsymbol{\delta}}^{(k)}) \\ &= \boldsymbol{Z}(\boldsymbol{I}_N - \eta\boldsymbol{D}_\eta^{-1/2}\boldsymbol{H}\boldsymbol{D}_\eta^{-1/2})\boldsymbol{Z}^\top\tilde{\boldsymbol{\delta}}^{(k)} + \tilde{\boldsymbol{q}}(\tilde{\boldsymbol{\delta}}^{(k)}),\end{aligned} \tag{5.55}$$

where $\tilde{\boldsymbol{q}}(\tilde{\boldsymbol{\delta}}^{(k)}) = (\boldsymbol{D}_\eta^{-1/2} \otimes \boldsymbol{I}_2)\hat{\boldsymbol{q}}((\boldsymbol{D}_\eta^{1/2} \otimes \boldsymbol{I}_2)\tilde{\boldsymbol{\delta}}^{(k)})$ satisfies

$$\begin{aligned}\left\|\tilde{\boldsymbol{q}}(\tilde{\boldsymbol{\delta}}^{(k)})\right\| &\le \left\|\boldsymbol{D}_\eta^{-1/2} \otimes \boldsymbol{I}_2\right\|_2 \left\|\hat{\boldsymbol{q}}((\boldsymbol{D}_\eta^{1/2} \otimes \boldsymbol{I}_2)\tilde{\boldsymbol{\delta}}^{(k)})\right\| \\ &= \left\|\boldsymbol{D}_\eta^{-1/2}\right\|_2 \left\|\hat{\boldsymbol{q}}\big((\boldsymbol{D}_\eta^{1/2} \otimes \boldsymbol{I}_2)\tilde{\boldsymbol{\delta}}^{(k)}\big)\right\| \\ &\le \left\|\boldsymbol{D}_\eta^{-1/2}\right\|_2 \cdot 2c_\eta(c_\eta+1)\left\|(\boldsymbol{D}_\eta^{1/2} \otimes \boldsymbol{I}_2)\tilde{\boldsymbol{\delta}}^{(k)}\right\|^2 \\ &\le 2c_\eta(c_\eta+1)\left\|\boldsymbol{D}_\eta^{-1/2}\right\|_2 \left\|\boldsymbol{D}_\eta^{1/2} \otimes \boldsymbol{I}_2\right\|_2^2 \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|^2 \\ &\le 2c_\eta(c_\eta+1)\left\|\boldsymbol{D}_\eta^{-1/2}\right\|_2 \left\|\boldsymbol{D}_\eta^{1/2}\right\|_2^2 \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|^2 \\ &= 2c_\eta(c_\eta+1)(1-\eta\underline{\gamma})^{1/2}(1-\eta\overline{\gamma})^{-1}\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|^2,\end{aligned}$$

where the last equality stems from $\|\boldsymbol{D}_\eta^{-1/2}\|_2 = (1-\eta\underline{\gamma})^{1/2}$ and $\|\boldsymbol{D}_\eta^{1/2}\|_2 = (1-\eta\overline{\gamma})^{-1/2}$. Let $q = 2c_\eta(c_\eta+1)(1-\eta\underline{\gamma})^{1/2}(1-\eta\overline{\gamma})^{-1}$. Taking the norm of both sides of (5.55) and then using the triangle inequality on the RHS, we obtain

$$\begin{aligned}\left\|\tilde{\boldsymbol{\delta}}^{(k+1)}\right\| &= \left\|\boldsymbol{Z}(\boldsymbol{I}_N - \eta\boldsymbol{D}_\eta^{-1/2}\boldsymbol{H}\boldsymbol{D}_\eta^{-1/2})\boldsymbol{Z}^\top\tilde{\boldsymbol{\delta}}^{(k)} + \tilde{\boldsymbol{q}}(\tilde{\boldsymbol{\delta}}^{(k)})\right\| \\ &\leq \left\|\boldsymbol{Z}(\boldsymbol{I}_N - \eta\boldsymbol{D}_\eta^{-1/2}\boldsymbol{H}\boldsymbol{D}_\eta^{-1/2})\boldsymbol{Z}^\top\tilde{\boldsymbol{\delta}}^{(k)}\right\| + \left\|\tilde{\boldsymbol{q}}(\tilde{\boldsymbol{\delta}}^{(k)})\right\|.\end{aligned}$$

Since $\boldsymbol{Z}(\boldsymbol{I}_N - \eta\boldsymbol{D}_\eta^{-1/2}\boldsymbol{H}\boldsymbol{D}_\eta^{-1/2})\boldsymbol{Z}^\top$ is symmetric, its spectral norm equals to its spectral radius. The last inequality can be rewritten as

$$\left\|\tilde{\boldsymbol{\delta}}^{(k+1)}\right\| \leq \rho\big(\boldsymbol{Z}(\boldsymbol{I}_N - \eta\boldsymbol{D}_\eta^{-1/2}\boldsymbol{H}\boldsymbol{D}_\eta^{-1/2})\boldsymbol{Z}^\top\big)\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| + q\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|^2. \tag{5.56}$$

Moreover, it can be seen from (5.55) that

$$\boldsymbol{Z}(\boldsymbol{I}_N - \eta\boldsymbol{D}_\eta^{-1/2}\boldsymbol{H}\boldsymbol{D}_\eta^{-1/2})\boldsymbol{Z}^\top = (\boldsymbol{D}_\eta^{-1/2}\otimes\boldsymbol{I}_2)\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top\big(\boldsymbol{D}_\eta^{-1/2}\otimes\boldsymbol{I}_2\big)^{-1},$$

which in turns implies the two matrices $\boldsymbol{Z}(\boldsymbol{I}_N - \eta\boldsymbol{D}_\eta^{-1/2}\boldsymbol{H}\boldsymbol{D}_\eta^{-1/2})\boldsymbol{Z}^\top$ and $\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top$ are similar and have the same spectral radius. In particular, we have

$$\begin{aligned}\rho\big(\boldsymbol{Z}(\boldsymbol{I}_N - \eta\boldsymbol{D}_\eta^{-1/2}\boldsymbol{H}\boldsymbol{D}_\eta^{-1/2})\boldsymbol{Z}^\top\big) &= \rho(\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top) \\ &= \rho(\boldsymbol{M}_\eta),\end{aligned}$$

where the second equality stems from Lemma 5.12. Thus, (5.56) can be represented as

$$\left\|\tilde{\boldsymbol{\delta}}^{(k+1)}\right\| \leq \rho(\boldsymbol{M}_\eta)\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| + q\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|^2.$$

Applying Lemma 5.14 with $b_k = \|\tilde{\boldsymbol{\delta}}^{(k)}\|$, $\rho = \rho(\boldsymbol{M}_\eta)$, and

$$c = \frac{\rho(\boldsymbol{M}_\eta)\big(1-\rho(\boldsymbol{M}_\eta)\big)}{q} = (1-\eta\underline{\gamma})^{1/2} c_1(\boldsymbol{x}^*, \eta),$$

it holds that if $\|\tilde{\boldsymbol{\delta}}^{(0)}\| < c$, then

$$\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| \leq \left(1 - \frac{\left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\|}{c}\right)^{-1} \left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\| \rho^k(\boldsymbol{M}_\eta). \tag{5.57}$$

Recall that $\boldsymbol{\delta}^{(k)} = (\boldsymbol{D}_\eta^{1/2} \otimes \boldsymbol{I}_2)\tilde{\boldsymbol{\delta}}^{(k)}$. On the one hand, the LHS of (5.57) can be lower-bounded by $(1-\eta\overline{\gamma})^{1/2}\left\|\boldsymbol{\delta}^{(k)}\right\|$ since

$$\begin{aligned}\left\|\boldsymbol{\delta}^{(k)}\right\| &= \left\|(\boldsymbol{D}_\eta^{1/2} \otimes \boldsymbol{I}_2)\tilde{\boldsymbol{\delta}}^{(k)}\right\| \leq \left\|\boldsymbol{D}_\eta^{1/2} \otimes \boldsymbol{I}_2\right\|_2 \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| \\ &= \left\|\boldsymbol{D}_\eta^{1/2}\right\|_2 \left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\| = (1-\eta\overline{\gamma})^{-1/2}\left\|\tilde{\boldsymbol{\delta}}^{(k)}\right\|.\end{aligned}$$

On the other hand, the RHS of (5.57) can be upper-bounded as follows. Since

$$\begin{aligned}\left\|\tilde{\boldsymbol{\delta}}^{(0)}\right\| &= \left\|(\boldsymbol{D}_\eta^{-1/2} \otimes \boldsymbol{I}_2)\boldsymbol{\delta}^{(0)}\right\| \leq \left\|\boldsymbol{D}_\eta^{-1/2} \otimes \boldsymbol{I}_2\right\|_2 \left\|\boldsymbol{\delta}^{(0)}\right\| \\ &= \left\|\boldsymbol{D}_\eta^{-1/2}\right\|_2 \left\|\boldsymbol{\delta}^{(0)}\right\| = (1-\eta\underline{\gamma})^{1/2}\left\|\boldsymbol{\delta}^{(0)}\right\|, \end{aligned} \tag{5.58}$$

we have

$$\left(1-\frac{\|\tilde{\boldsymbol{\delta}}^{(0)}\|}{c}\right)^{-1}\|\tilde{\boldsymbol{\delta}}^{(0)}\|\rho^k(\boldsymbol{M}_\eta)\leq\left(1-\frac{(1-\eta\underline{\gamma})^{1/2}\|\boldsymbol{\delta}^{(0)}\|}{c}\right)^{-1}(1-\eta\underline{\gamma})^{1/2}\|\boldsymbol{\delta}^{(0)}\|\rho^k(\boldsymbol{M}_\eta)$$
$$=\left(1-\frac{\|\boldsymbol{\delta}^{(0)}\|}{c_1(\boldsymbol{x}^*,\eta)}\right)^{-1}(1-\eta\underline{\gamma})^{1/2}\|\boldsymbol{\delta}^{(0)}\|\rho^k(\boldsymbol{M}_\eta). \tag{5.59}$$

From the lower bound $(1-\eta\overline{\gamma})^{1/2}\left\|\boldsymbol{\delta}^{(k)}\right\|$ and the upper bound in (5.59), we obtain (5.26). Finally, the region of convergence $\|\boldsymbol{\delta}^{(0)}\|<c_1(\boldsymbol{x}^*,\eta)$ is sufficient to guarantee that $\|\tilde{\boldsymbol{\delta}}^{(0)}\|<c=(1-\eta\underline{\gamma})^{1/2}c_1(\boldsymbol{x}^*,\eta)$ due to (5.58). This completes our proof of the lemma.

5.8.11 Auxiliary Lemmas

Lemma 5.11. *Given a matrix $\boldsymbol{Z}\in\mathbb{R}^{2N\times N}$ as in (5.15). Then for any diagonal matrix $\boldsymbol{D}\in\mathbb{R}^{N\times N}$, we have $(\boldsymbol{D}\otimes\boldsymbol{I}_2)\boldsymbol{Z}=\boldsymbol{Z}\boldsymbol{D}$.*

Proof. Recall from (5.46) that $\boldsymbol{Z}=\sum_{i=1}^{N}\boldsymbol{e}_i\boldsymbol{e}_i^\top\otimes\boldsymbol{v}_i$, where $\boldsymbol{v}_i=[-x_{2i},x_{2i-1}]^\top$. By

representing $\boldsymbol{D} = \sum_{i=1}^{N} D_{ii}\boldsymbol{e}_i\boldsymbol{e}_i^\top$, we have

$$\begin{aligned}
(\boldsymbol{D}\otimes\boldsymbol{I}_2)\boldsymbol{Z} &= \Big((\sum_{i=1}^{N} D_{ii}\boldsymbol{e}_i\boldsymbol{e}_i^\top)\otimes\boldsymbol{I}_2\Big)\cdot(\sum_{j=1}^{N}\boldsymbol{e}_j\boldsymbol{e}_j^\top\otimes\boldsymbol{v}_j)\\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\Big(\big(D_{ii}\boldsymbol{e}_i\boldsymbol{e}_i^\top\big)\cdot\big(\boldsymbol{e}_j\boldsymbol{e}_j^\top\big)\Big)\otimes(\boldsymbol{I}_2\cdot\boldsymbol{v}_j)\\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} D_{ii}\big((\boldsymbol{e}_i^\top\boldsymbol{e}_j)\cdot\boldsymbol{e}_i\boldsymbol{e}_j^\top\big)\otimes\boldsymbol{v}_j\\
&= \sum_{i=1}^{N} D_{ii}(\boldsymbol{e}_i\boldsymbol{e}_i^\top)\otimes\boldsymbol{v}_i\\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\Big(\big(\boldsymbol{e}_i\boldsymbol{e}_i^\top\big)\cdot\big(D_{jj}\boldsymbol{e}_j\boldsymbol{e}_j^\top\big)\Big)\otimes(\boldsymbol{v}_i\cdot 1)\\
&= \Big(\sum_{i=1}^{N}\boldsymbol{e}_i\boldsymbol{e}_i^\top\otimes\boldsymbol{v}_i\Big)\cdot\Big((\sum_{i=1}^{N} D_{jj}\boldsymbol{e}_j\boldsymbol{e}_j^\top)\otimes 1\Big)\\
&= \Big(\sum_{i=1}^{N}\boldsymbol{e}_i\boldsymbol{e}_i^\top\otimes\boldsymbol{v}_i\Big)\cdot(\sum_{i=1}^{N} D_{jj}\boldsymbol{e}_j\boldsymbol{e}_j^\top)\\
&= \boldsymbol{Z}\boldsymbol{D},
\end{aligned}$$

where it is noted that

$$\boldsymbol{e}_i^\top\boldsymbol{e}_j = \begin{cases} 1 & \text{if } i = j,\\ 0 & \text{if } i \neq j. \end{cases}$$

□

Lemma 5.12. *For any eigenvalue λ of $\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top$, either $\lambda = 0$ or λ is an eigenvalue*

of $\boldsymbol{M}_\eta$*. Consequently, we have*

$$\rho(\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top) = \rho(\boldsymbol{M}_\eta).$$

Proof. Let $(\lambda, \boldsymbol{u})$ be a pair of eigenvalue and eigenvector of $\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top$. Then, we have

$$\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top\boldsymbol{u} = \lambda\boldsymbol{u}. \tag{5.60}$$

Left-multiplying both sides of (5.60) by $\boldsymbol{Z}^\top$ and using the semi-orthogonality of $\boldsymbol{Z}$, we obtain

$$\boldsymbol{M}_\eta(\boldsymbol{Z}^\top\boldsymbol{u}) = \lambda(\boldsymbol{Z}^\top\boldsymbol{u}).$$

This means either $\boldsymbol{Z}^\top\boldsymbol{u} = \boldsymbol{0}_N$ or $\boldsymbol{Z}^\top\boldsymbol{u}$ is an eigenvector of $\boldsymbol{M}_\eta$. In the former case, we have $\lambda = 0$. In the latter case, we have λ is an eigenvalue of $\boldsymbol{M}_\eta$. Finally, since the spectral radius is the maximum absolute value of all eigenvalues, it is trivial that $\rho(\boldsymbol{Z}\boldsymbol{M}_\eta\boldsymbol{Z}^\top) = \rho(\boldsymbol{M}_\eta)$. □

Lemma 5.13. *(Rephrased from the supplemental material of [212]) Let* $\{a_k\}_{k=0}^\infty \subset \mathbb{R}_+$ *be the sequence defined by*

$$a_{k+1} = \rho a_k + q a_k^2 \qquad \text{for } k = 0, 1, \ldots, \tag{5.61}$$

where $0 < \rho < 1$ *and* $q \geq 0$*. Then* $\{a_k\}_{k=0}^\infty$ *converges* ***monotonically*** *to* 0 *if and*

only if $a_0 < \frac{1-\rho}{q}$. *A simple linear convergence bound can be derived for* $a_0 < \rho\frac{1-\rho}{q}$ *in the form of*

$$a_k \leq \left(1 - \frac{a_0 q}{\rho(1-\rho)}\right)^{-1} a_0 \rho^k. \tag{5.62}$$

Proof. For each $k \in \mathbb{N}$, let us define $d_k = a_k/(a_0\rho^k)$. Substituting $a_k = a_0 d_k \rho^k$ into (5.61) and defining $\tau = a_0 q/(1-\rho)$, we obtain

$$\begin{cases} d_0 = 1, \\ d_{k+1} = d_k + \tau(1-\rho)\rho^{k-1} d_k^2 \quad \text{for } k = 0, 1, \ldots. \end{cases}$$

Since $\tau(1-\rho)\rho^{k-1}d_k^2 > 0$, the sequence $\{d_k\}_{k=0}^{\infty}$ is strictly increasing and positive. Thus, using $d_i > d_{i+1} > 0$, for any $i = 0, 1, \ldots, k-1$, we have

$$\frac{1}{d_i} - \frac{1}{d_{i+1}} = \frac{d_{i+1} - d_i}{d_{i+1} d_i} < \frac{d_{i+1} - d_i}{d_i^2} = \tau(1-\rho)\rho^{i-1}.$$

Summing over $i = 0, 1, \ldots, k-1$, we obtain

$$1 - \frac{1}{d_k} < \sum_{i=0}^{k-1} \tau(1-\rho)\rho^{i-1} = \frac{\tau}{\rho}(1-\rho^k) < \frac{\tau}{\rho}. \tag{5.63}$$

Substituting $d_k = a_k/(a_0\rho^k)$ and $\tau = a_0 q/(1-\rho)$ into (5.63) and rearranging terms yield the desired bound on a_k in (5.62). □

Lemma 5.14. *Let* $\{b_k\}_{k=0}^{\infty} \subset \mathbb{R}_+$ *be the sequence defined by*

$$b_{k+1} \leq \rho b_k + q b_k^2 \qquad \text{for } k = 0, 1, \ldots, \tag{5.64}$$

where $0 < \rho < 1$ *and* $q \geq 0$. *If* $b_0 < \frac{1-\rho}{q}$, *then* $\{b_k\}_{k=0}^{\infty}$ *converges to* 0. *If* $b_0 < c \triangleq \rho \frac{1-\rho}{q}$, *then for any integer* $k \geq 0$, *we have*

$$b_k \leq \left(1 - \frac{b_0}{c}\right)^{-1} b_0 \rho^k.$$

Proof. Let us define a surrogate sequence $\{a_k\}_{k=0}^{\infty}$ that upper-bounds $\{b_k\}_{k=0}^{\infty}$ as follows

$$\begin{cases} a_0 = b_0, \\ a_{k+1} = \rho a_k + q a_k^2. \end{cases}$$

First, we prove by induction that

$$b_k \leq a_k \quad \forall k \in \mathbb{N}. \tag{5.65}$$

The base case when $k = 0$ holds trivially as $b_0 = a_0$. In the induction step, given $b_k \leq a_k$ for an integer $k \geq 0$, we have

$$b_{k+1} \leq \rho b_k + q b_k^2 \leq \rho a_k + a_k^2 = a_{k+1}.$$

By the principle of induction, (5.65) holds for all $k \in \mathbb{N}$. Now, by Lemma 5.13, we

have

$$\begin{aligned} b_k \leq a_k &\leq \left(1 - \frac{a_0 q}{\rho(1-\rho)}\right)^{-1} a_0 \rho^k \\ &= \left(1 - \frac{b_0 q}{\rho(1-\rho)}\right)^{-1} b_0 \rho^k. \end{aligned}$$

This completes our proof of the lemma. □

Chapter 6: Perturbation Expansions and Error Bounds for the Truncated Singular Value Decomposition[1]

Truncated singular value decomposition is a reduced version of the singular value decomposition in which only a few largest singular values are retained. This chapter presents a novel perturbation analysis for the truncated singular value decomposition for real matrices. First, we describe perturbation expansions for the singular value truncation of order r. We extend perturbation results for the singular subspace decomposition to derive the first-order perturbation expansion of the truncated operator about a matrix with rank greater than or equal to r. Observing that the first-order expansion can be greatly simplified when the matrix has exact rank r, we further show that the singular value truncation admits a simple second-order perturbation expansion about a rank-r matrix. Second, we introduce the first-known error bound on the linear approximation of the truncated singular value decomposition of a perturbed rank-r matrix. Our bound only depends on the least singular value of the unperturbed matrix and the norm of the perturbation matrix. Intriguingly, while the singular subspaces are known to be extremely sensitive to additive noises, the newly established error bound holds universally

[1]This work has been published as: Trung Vu, Evgenia Chunikhina, and Raviv Raich. "Perturbation expansions and error bounds for the truncated singular value decomposition." Linear Algebra and its Applications, vol. 627, pp. 94-139, 2021.

for perturbations with arbitrary magnitude. Finally, we demonstrate an application of our results to the analysis of the mean squared error associated with the TSVD-based matrix denoising solution.

6.1 Introduction

The singular value decomposition (SVD) is an invaluable tool for matrix analysis and the truncated singular value decomposition (TSVD) offers a formal approach for a rank-restricted optimal approximation of matrices by replacing the smallest singular values by zeros in the SVD of a matrix. TSVD has numerous applications in science, engineering, and math with examples including linear system identification [146,147], collaborative filtering [33,104], low-rank matrix denoising [182,231], data compression [225], and numerical partial differential equations [131]. In addition, TSVD is well-known for solving classical discrete ill-posed problems [89, 90]. This chapter is concerned with the effects of errors on the truncated singular value decomposition of a matrix.

Perturbation theory for the SVD studies the effect of variation in matrix entries on the singular values and the singular vectors of a matrix. Using perturbation bounds or perturbation expansions, one can characterize the difference between the SVD-related quantities associated with the perturbed matrix and those of the original matrix. The first perturbation bound on singular values was given by Weyl [226] in 1912, stating that no singular value can be changed by more than the spectral norm of the perturbation. Later, Mirsky [153] showed that Weyl's

inequality also holds for any unitarily-invariant norm. Perturbation bounds for singular vectors are often established in the context of singular subspace decomposition. In 1970, Davis and Kahan [52] introduced a fundamental bound on the distance between the subspaces spanned by a group of eigenvectors and their perturbed versions based the ratio between the perturbation level and the eigengap. This result is also referred as the so-called $\sin\Theta$ theorem for symmetric matrices in the literature. Shortly afterwards, Wedin [224] generalized part of this result to cover non-symmetric matrices using the singular value decomposition, bounding changes in the left and right singular subspaces in terms of the singular value gap and the perturbation magnitude. In a recent work, Cai and Zhang [29] further established separate matching upper and lower bounds for the left and right singular subspaces. When the structure of the error is concerned, one may draw interest in perturbation expansions to approximate the perturbed quantity as a function of the perturbation matrix. As the perturbation decreases towards zero, the approximation is more accurate since the higher-order terms in the expansion become successively smaller. In 1973, Stewart [190] showed that there exists explicit expression of the perturbed subspaces in the bases of the unperturbed subspaces, which can be leveraged to obtain error bounds for certain characteristic subspaces associated with the SVD. This breakthrough result has started a long line of research on perturbation expansions and error bounds for the SVD, including the work of Stewart [191], Sun [198], Li *et al.* [130], Vaccaro [208], Xu [230], Liu *et al.* [135], and more recently, Gratton *et al.* [82]. Specifically, in [191], Stewart utilized the bounding technique in [190] and obtained a second-order perturbation expansion

for the square of the smallest singular value of a matrix. In a different approach based on the theory of implicit functions, Sun [198] provided the first analytical expression for the second-order perturbation expansion of simple non-zero singular values of a matrix. One of the first significant results on perturbation expansion of singular subspaces was introduced by Li and Vaccaro in 1991. In [130], the two authors analyzed a variety of subspace-based algorithms in array signal processing and developed the first-order perturbation expansion for the signal and orthogonal subspaces of the rank-deficient data matrix. Later on, Vaccaro [208] extended this result to the second-order perturbation expansion of these subspaces. A more fine-grained analysis of the perturbation expansion for the individual singular vectors rather than the singular subspaces was given by Liu *et al.* [135], uncovering the fact that the signal subspace has an impact on the first-order approximation of the individual singular vectors, but not on the first-order approximation of the signal subspace spanned by these vectors. We note that the aforementioned results on perturbation analysis of singular subspaces make an assumption that the unperturbed matrix is rank-deficient, i.e., all singular values corresponding to one of the singular subspaces are zero. In 2002, Xu [230] relaxed this constraint by only requiring those singular values to be equally small. Recently, Gratton and Tshimanga [82] were able to eliminate this constraint completely, presenting the second-order perturbation expansion for singular subspaces with no restriction on their corresponding singular values.[2] It is notable that the last result is developed directly from those by Stewart in [190]. A more comprehensive description

[2]The only constraint is the singular-value separation between the two subspaces.

of the aforementioned results is given in Section 3. Interested readers can also find in-depth surveys on matrix perturbation theory in [192, 193] and references therein.

The aforementioned results on perturbation analysis of the SVD is the fulcrum for the perturbation analysis of the TSVD. While the former characterizes the effect of perturbation on the singular values/singular subspaces of a matrix, the later studies the combined effect (from both singular values and singular subspaces) on the resulting reduced-rank matrix. Analyzing such an effect helps understand the local behavior of algorithms that utilize the low-rank optimal approximation of matrices, such as SVD-based channel estimation methods in multi-input multi-output (MIMO) systems [109, 134, 162] and iterative hard-thresholding algorithms for low-rank matrix completion [104, 213, 214]. In a recent work, Gratton and Tshimanga [82] presented a second-order expansion for the singular subspace decomposition and make use of the result to deduce the second-order sensitivity of the TSVD solution to least-squares problems. However, since their application focuses on the expansion of the truncated pseudo-inverse rather than the TSVD itself, no specific result in perturbation expansion of the TSVD is mentioned. In a different approach to analyzing the TSVD operator, Feppon and Lermusiaux [64] studied the embedded geometry of the fixed-rank matrix manifold and characterized the projection onto it as a smooth (C^∞) map. Based on this geometric interpretation, the authors provided an explicit expression for the directional derivative of the TSVD of order r at a certain matrix with rank greater than or equal to r.[3] On

[3]Despite the fact that Theorem 25 in [64] reads "greater than r", both the proof of the theorem

the one hand, the result directly suggests the first-order perturbation expansion of the TSVD. On the other hand, the differential geometry-based approach, while offering a clear path for calculating the derivatives, does not offer a direct recipe for obtaining the error bound on the first-order approximation or the higher order terms in the expansion. At the time of writing this chapter, we are not aware of any explicit expression of the second-order derivative of the TSVD.

In this chapter, we present a novel perturbation analysis of the truncated singular value decomposition. First, by utilizing the perturbation expansion for singular subspaces in [82], we derive the first-order perturbation expansion of the TSVD. Our result matches the result on the directional derivative of the TSVD in [64]. Furthermore, we extend our analysis to study the second-order perturbation expansion and show that when the matrix has exact rank r, the TSVD of order r admits a simple expression for its second-order expansion. To the best of our knowledge, this is the first explicit result for the second-order perturbation expansion of the TSVD. Third, we establish an error bound on the first-order approximation of the TSVD about a rank-r matrix. Our bound holds universally for any level (or magnitude) of the perturbation. Finally, we demonstrate how the proposed perturbation expansions and error bounds can be applied to study the mean squared error associated with the TSVD-based matrix denoising solution.

---

and the direct communication with the authors (on September 17, 2020) suggest the result should also include the case of rank-r matrices.

6.2 Notation and Definitions

Throughout the chapter, we use $\|\cdot\|_F$ and $\|\cdot\|_2$ to denote the Frobenius norm and the spectral norm of a matrix, respectively. Occasionally, $\|\cdot\|_2$ is used on a vector to denote the Euclidean norm. Boldfaced symbols are reserved for vectors and matrices. In addition, the $s \times t$ all-zero matrix is denoted by $\mathbf{0}_{s\times t}$ and the $s \times s$ identity matrix is denoted by $\boldsymbol{I}_s$. We also use $\boldsymbol{e}_i^s$ to denote the i-th vector in the natural basis of $\mathbb{R}^s$. When understood clearly from the context, the dimensions of vectors/matrices in the aforementioned notation may be omitted. As a slight abuse of notation, we define the big O notation for matrices as follows.

Definition 6.1. *Let $\boldsymbol{\Delta}$ be some matrix and $\boldsymbol{F}(\boldsymbol{\Delta})$ be a matrix-valued function of $\boldsymbol{\Delta}$. Then, for any positive number k, $\boldsymbol{F}(\boldsymbol{\Delta}) = \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^k)$ if there exists some constant $0 \leq c < \infty$ such that*

$$\lim_{\epsilon\to 0^+} \sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \frac{\|\boldsymbol{F}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^k} = c.$$

We emphasize the difference between the commonly used big O notation in the literature and the $\boldsymbol{\mathcal{O}}$ notation used in this chapter. While the former requires c to be strictly greater than 0, our notation includes the case $c = 0$ to imply both situations that $\boldsymbol{F}(\boldsymbol{\Delta})$ approaches $\mathbf{0}$ at a rate either equal or faster than $\|\boldsymbol{\Delta}\|_F^k$. Similarly, when used for a vector, we replace the Frobenius norm by the Euclidean norm in Definition 6.1 to denote the corresponding quantity.

In the rest of the chapter, unless otherwise specified, the symbol $\boldsymbol{X}$ is used to denote an arbitrary matrix in $\mathbb{R}^{m\times n}$. Here, without loss of generality, we assume

that $m \geq n$. The SVD of $\boldsymbol{X}$ is written as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ where $\boldsymbol{\Sigma}$ is a $m \times n$ rectangular diagonal matrix with main diagonal entries are the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$. For completeness, we denote the "ghost" singular values $\sigma_{n+1} = \ldots = \sigma_m = 0$ in the case $m > n$. Additionally, $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices such that $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_m$ and $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_n$. We note that the left and right singular vectors of $\boldsymbol{X}$ are the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, i.e., $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$. Thus, $\boldsymbol{X}$ can also be rewritten as the sum of rank-1 matrices: $\boldsymbol{X} = \sum_{i=1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$. Next, we define the singular subspace decomposition as follows.

Definition 6.2. *Given $1 \leq r < n$, the singular subspace decomposition of $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is given by:*

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^T \\ \boldsymbol{V}_2^T \end{bmatrix} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T + \boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T, \tag{6.1}$$

where

$$\boldsymbol{\Sigma}_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}, \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} \mathrm{diag}(\sigma_{r+1}, \ldots, \sigma_n) \\ \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{(m-r)\times(n-r)},$$

with the singular values in descending order, i.e., $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$, and

$$\boldsymbol{U}_1 = \begin{bmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_r \end{bmatrix} \in \mathbb{R}^{m\times r}, \quad \boldsymbol{U}_2 = \begin{bmatrix} \boldsymbol{u}_{r+1} & \ldots & \boldsymbol{u}_m \end{bmatrix} \in \mathbb{R}^{m\times(m-r)},$$
$$\boldsymbol{V}_1 = \begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_r \end{bmatrix} \in \mathbb{R}^{n\times r}, \quad \boldsymbol{V}_2 = \begin{bmatrix} \boldsymbol{v}_{r+1} & \ldots & \boldsymbol{v}_n \end{bmatrix} \in \mathbb{R}^{n\times(n-r)}.$$

It is clear from Definition 6.2 that

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 \end{bmatrix}, \qquad \boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 \end{bmatrix}.$$

Here the columns of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ (or $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$) provide the bases for the column-space (or row-space) of $\boldsymbol{X}$ and its orthogonal complement, respectively.

Definition 6.3. *The orthonormal projectors onto the subspaces of $\boldsymbol{X}$ are defined as:*

$$\boldsymbol{P}_{\boldsymbol{U}_1} = \boldsymbol{U}_1\boldsymbol{U}_1^T = \sum_{i=1}^{r} \boldsymbol{u}_i\boldsymbol{u}_i^T, \qquad \boldsymbol{P}_{\boldsymbol{U}_2} = \boldsymbol{U}_2\boldsymbol{U}_2^T = \boldsymbol{I}_m - \boldsymbol{P}_{\boldsymbol{U}_1} = \sum_{i=r+1}^{m} \boldsymbol{u}_i\boldsymbol{u}_i^T,$$
$$\boldsymbol{P}_{\boldsymbol{V}_1} = \boldsymbol{V}_1\boldsymbol{V}_1^T = \sum_{i=1}^{r} \boldsymbol{v}_i\boldsymbol{v}_i^T, \qquad \boldsymbol{P}_{\boldsymbol{V}_2} = \boldsymbol{V}_2\boldsymbol{V}_2^T = \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{V}_1} = \sum_{i=r+1}^{n} \boldsymbol{v}_i\boldsymbol{v}_i^T.$$

Generally, matrices $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{V}_1$, and $\boldsymbol{V}_2$ are not unique. In particular, for simple non-zero singular values, the corresponding left and right singular vectors are unique up to a simultaneous sign change. For repeated and positive singular values, the corresponding left and right singular vectors are unique up to a simultaneous right multiplication with the same orthogonal matrix. Finally, for zero singular values, the singular vectors can be any orthonormal bases of the left and right null spaces of $\boldsymbol{X}$. On the other hand, the singular subspaces spanned by the columns of $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{V}_1$, $\boldsymbol{V}_2$, and their corresponding projectors are unique provided that $\sigma_r > \sigma_{r+1}$ [89]. We are now in position to define the singular value truncation.

Definition 6.4. *The r-truncated singular value decomposition of $\boldsymbol{X}$ (r-TSVD) is*

defined as

$$\mathcal{P}_r(\boldsymbol{X}) = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T = \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^T. \tag{6.2}$$

By Eckart-Young theorem [61], $\mathcal{P}_r(\boldsymbol{X})$ is the best least squares approximation of $\boldsymbol{X}$ by a rank-r matrix, with respect to unitarily-invariant norms. Therefore, this operator is also known as the projection of $\boldsymbol{X}$ onto the non-convex set of rank-r matrices. $\mathcal{P}_r(\boldsymbol{X})$ is unique if either $\sigma_r > \sigma_{r+1}$ or $\sigma_r = 0$. In the special case when $\boldsymbol{X}$ has exact rank r, we have $\sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$ and the projectors onto the subspaces of $\boldsymbol{X}$, namely, $\boldsymbol{P}_{\boldsymbol{U}_1}, \boldsymbol{P}_{\boldsymbol{U}_2}, \boldsymbol{P}_{\boldsymbol{V}_1}$, and $\boldsymbol{P}_{\boldsymbol{V}_2}$ are unique. However, the matrices $\boldsymbol{U}_2$ and $\boldsymbol{V}_2$ can take any orthonormal basis in $\boldsymbol{R}^{m-r}$ and $\boldsymbol{R}^{n-r}$, respectively, as their columns. Finally, for a rank-r matrix, we define the pseudo inverse of $\boldsymbol{X}$ as $\boldsymbol{X}^\dagger = \boldsymbol{U}_1 \boldsymbol{\Sigma}_1^{-1} \boldsymbol{V}_1^T$. It is worth mentioning that $\|\boldsymbol{X}\|_2 = \sigma_1$ while $\left\|\boldsymbol{X}^\dagger\right\|_2 = 1/\sigma_r$ in this case.

6.3 Preliminaries

Two elemental bounds for singular values were given by Weyl [226] in 1912 and Mirsky [153] in 1960:

Proposition 6.1. *Let $\boldsymbol{\Delta} \in \mathbb{R}^{m \times n}$ be a perturbation of arbitrary magnitude. Denote $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta}$ with singular values $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \ldots \geq \tilde{\sigma}_n \geq 0$. Then,*

- *Weyl's inequality: $|\tilde{\sigma}_i - \sigma_i| \leq \|\boldsymbol{\Delta}\|_2$, for $i = 1, \ldots, n$,*
- *Mirsky's inequality: $\sqrt{\sum_{i=1}^{n} (\tilde{\sigma}_i - \sigma_i)^2} \leq \|\boldsymbol{\Delta}\|_F$.*

Proposition 6.1 asserts that the changes in the singular values can be bounded using only the norm of the perturbation. By leveraging the specific values of the entries of the perturbation matrix, the behavior of singular values under perturbations can be described more precisely through perturbation expansions. In [191], Stewart showed that if σ_n is non-zero and distinct from other singular values of $\boldsymbol{X}$, then its corresponding perturbed singular value can be expressed by

$$\tilde{\sigma}_n = \sigma_n + \boldsymbol{u}_n^T \boldsymbol{\Delta} \boldsymbol{v}_n + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|^2). \tag{6.3}$$

It is later known that the result in (6.3) also holds for any simple non-zero singular values [193]. In another approach, Sun [198] derived a second-order perturbation expansion for simple non-zero singular values. For a simple zero singular value, Stewart [191] claimed that deriving a perturbation expansion is non-trivial and proposed a second-order approximation for $\tilde{\sigma}_n^2$ instead. Most recently, a generalization of (6.3) to a set of singular values that is well separated from the rest is proved in [194].

While the singular values of a matrix are proven to be quite stable under perturbations, the singular vectors, especially those correspond to a cluster of singular values, are extremely sensitive. It is therefore natural to bound the perturbation error based on the subspace spanned by the singular vectors. Consider the singular subspace decomposition in Definition 6.2. We define the singular gap as the smallest distance between a singular value in $\boldsymbol{\Sigma}_1$ and a singular value in $\boldsymbol{\Sigma}_2$. When the spectral norm of the perturbation is smaller than this gap, Wedin's

$\sin\Theta$ theorem [224] provides an upper bound on the distances between the left and right singular subspaces and their corresponding perturbed counterparts in terms of the singular gap and the Frobenius norm of the perturbation. Furthermore, Stewart [190] showed that there exist explicit expressions of the perturbed subspaces in the bases of the unperturbed subspaces, which can be leveraged to obtain error bounds for certain characteristic subspaces associated with the SVD. Let us rephrase this result in the following proposition.

Proposition 6.2. *(Rephrased from Theorem 2.1 in [82], which is based on Theorem 6.4 in [190]) In addition to the setting in Definition 6.2, assume that* $\sigma_r > \sigma_{r+1}$. *For a perturbation* $\boldsymbol{\Delta} \in \mathbb{R}^{m\times n}$, *denote the singular subspace decomposition of* $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta}$ *by*

$$\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^T = \begin{bmatrix}\tilde{\boldsymbol{U}}_1 & \tilde{\boldsymbol{U}}_2\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{\Sigma}}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \tilde{\boldsymbol{\Sigma}}_2\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{V}}_1^T\\ \tilde{\boldsymbol{V}}_2^T\end{bmatrix}.$$

Let us partition $\boldsymbol{U}^T\boldsymbol{\Delta}\boldsymbol{V}$ *conformally with* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *in the form*

$$\boldsymbol{U}^T\boldsymbol{\Delta}\boldsymbol{V} = \begin{bmatrix}\boldsymbol{U}_1^T\boldsymbol{\Delta}\boldsymbol{V}_1 & \boldsymbol{U}_1^T\boldsymbol{\Delta}\boldsymbol{V}_2\\ \boldsymbol{U}_2^T\boldsymbol{\Delta}\boldsymbol{V}_1 & \boldsymbol{U}_2^T\boldsymbol{\Delta}\boldsymbol{V}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{E}_{11} & \boldsymbol{E}_{12}\\ \boldsymbol{E}_{21} & \boldsymbol{E}_{22}\end{bmatrix} = \boldsymbol{E}. \tag{6.4}$$

If

$$\|\boldsymbol{\Delta}\|_2 < \frac{\sigma_r - \sigma_{r+1}}{2}, \tag{6.5}$$

then there must exist unique matrices $\boldsymbol{Q} \in \mathbb{R}^{(m-r)\times r}$, $\boldsymbol{P} \in \mathbb{R}^{(n-r)\times r}$ *whose norms are in the order of* $\|\boldsymbol{\Delta}\|_F$ *such that*

$$\boldsymbol{Q}(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11}) + (\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22})\boldsymbol{P} = -\boldsymbol{E}_{21} - \boldsymbol{Q}\boldsymbol{E}_{12}\boldsymbol{P}, \tag{6.6a}$$

$$(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})\boldsymbol{P}^T + \boldsymbol{Q}^T(\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22}) = \boldsymbol{E}_{12} + \boldsymbol{Q}^T\boldsymbol{E}_{21}\boldsymbol{P}^T. \tag{6.6b}$$

Moreover, using

$$\hat{\boldsymbol{U}}_1 = (\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{I}_r + \boldsymbol{Q}^T\boldsymbol{Q})^{-1/2}, \tag{6.7a}$$

$$\hat{\boldsymbol{U}}_2 = (\boldsymbol{U}_2 + \boldsymbol{U}_1\boldsymbol{Q}^T)(\boldsymbol{I}_{m-r} + \boldsymbol{Q}\boldsymbol{Q}^T)^{-1/2}, \tag{6.7b}$$

$$\hat{\boldsymbol{V}}_1 = (\boldsymbol{V}_1 + \boldsymbol{V}_2\boldsymbol{P})(\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1/2}, \tag{6.7c}$$

$$\hat{\boldsymbol{V}}_2 = (\boldsymbol{V}_2 - \boldsymbol{V}_1\boldsymbol{P}^T)(\boldsymbol{I}_{n-r} + \boldsymbol{P}\boldsymbol{P}^T)^{-1/2}, \tag{6.7d}$$

we can define semi-orthogonal matrices $\hat{\boldsymbol{U}}_1$, $\hat{\boldsymbol{U}}_2$, $\hat{\boldsymbol{V}}_1$, *and* $\hat{\boldsymbol{V}}_2$ *satisfying* $\hat{\boldsymbol{U}}_1^T\hat{\boldsymbol{U}}_2 = \boldsymbol{0}$ *and* $\hat{\boldsymbol{V}}_1^T\hat{\boldsymbol{V}}_2 = \boldsymbol{0}$, *which provide bases to the same unique subspaces of* $\tilde{\boldsymbol{U}}_1$, $\tilde{\boldsymbol{U}}_2$, $\tilde{\boldsymbol{V}}_1$, *and* $\tilde{\boldsymbol{V}}_2$, *respectively, i.e.,* $\boldsymbol{P}_{\hat{\boldsymbol{U}}_1} = \boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}$, $\boldsymbol{P}_{\hat{\boldsymbol{U}}_2} = \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}$, $\boldsymbol{P}_{\hat{\boldsymbol{V}}_1} = \boldsymbol{P}_{\tilde{\boldsymbol{V}}_1}$, *and* $\boldsymbol{P}_{\hat{\boldsymbol{V}}_2} = \boldsymbol{P}_{\tilde{\boldsymbol{V}}_2}$.

It is important to note that $\hat{\boldsymbol{U}}_1$, $\hat{\boldsymbol{U}}_2$, $\hat{\boldsymbol{V}}_1$, and $\hat{\boldsymbol{V}}_2$ may differ from $\tilde{\boldsymbol{U}}_1$, $\tilde{\boldsymbol{U}}_2$, $\tilde{\boldsymbol{V}}_1$, and $\tilde{\boldsymbol{V}}_2$, respectively. However, their corresponding subspaces are identical. This result will be useful later when replacing $\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}$ and $\boldsymbol{P}_{\tilde{\boldsymbol{V}}_1}$ in the following version of the r-TSVD $\mathcal{P}_r(\tilde{\boldsymbol{X}}) = \boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_1}$ with $\boldsymbol{P}_{\hat{\boldsymbol{U}}_1}$ and $\boldsymbol{P}_{\hat{\boldsymbol{V}}_1}$. The substitution allows us to write an explicit expression of the r-TSVD using $\boldsymbol{\Delta}$ and terms that are in order of $\|\boldsymbol{\Delta}\|_F$ such as $\boldsymbol{Q}$ and $\boldsymbol{P}$. Equation (6.6) also enables the perturbation expansion of the SVD through the coefficient matrices $\boldsymbol{Q}$ and $\boldsymbol{P}$. In 1991, Li and

Vaccaro [130] considered a special case of rank-r matrices ($\boldsymbol{\Sigma}_2 = \mathbf{0}$) and introduced the first-order perturbation expansion for $\boldsymbol{Q}$ and $\boldsymbol{P}$ as a method to analyze the performance of subspace-based algorithms in array signal processing. Later on, Vaccaro [208] extended their approach to study the second-order perturbation expansion for the singular subspace decomposition. A more general result in this approach was proposed by Xu [230] in 2002, through relaxing the constraint $\boldsymbol{\Sigma}_2 = \mathbf{0}$ to $\boldsymbol{\Sigma}_2^T\boldsymbol{\Sigma}_2 = \epsilon^2\boldsymbol{I}$, for small $\epsilon \geq 0$. It was not until recently the second-order analysis with no restriction on $\boldsymbol{\Sigma}_2$ was provided by Gratton [82]. We summarize this result on second-order perturbation expansion for $\boldsymbol{Q}$ and $\boldsymbol{P}$ as follows.

Proposition 6.3. *Given the setting in Proposition 6.2. Then*

$$\operatorname{vec}(\boldsymbol{Q}) = \boldsymbol{\Phi}_0^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Phi}_0^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Phi}_0^{-1}\boldsymbol{\Phi}_1\boldsymbol{\Phi}_0^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^3), \tag{6.8}$$

where

$$\begin{aligned}
&\boldsymbol{\Phi}_0 = \boldsymbol{\Sigma}_1^2 \otimes \boldsymbol{I}_{m-r} - \boldsymbol{I}_r \otimes (\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^T),\\
&\boldsymbol{\Phi}_1 = (\boldsymbol{\Sigma}_1\boldsymbol{E}_{11}^T + \boldsymbol{E}_{11}\boldsymbol{\Sigma}_1) \otimes \boldsymbol{I}_{m-r} - \boldsymbol{I}_r \otimes (\boldsymbol{\Sigma}_2\boldsymbol{E}_{22}^T + \boldsymbol{E}_{22}\boldsymbol{\Sigma}_2^T),\\
&\boldsymbol{\mu}_1 = -\operatorname{vec}(\boldsymbol{\Sigma}_2\boldsymbol{E}_{12}^T + \boldsymbol{E}_{21}\boldsymbol{\Sigma}_1), \qquad \boldsymbol{\mu}_2 = -\operatorname{vec}(\boldsymbol{E}_{22}\boldsymbol{E}_{12}^T + \boldsymbol{E}_{21}\boldsymbol{E}_{11}^T),
\end{aligned}$$

and

$$\operatorname{vec}(\boldsymbol{P}) = \boldsymbol{\Psi}^{-1}\boldsymbol{\tau}_1 + \boldsymbol{\Psi}_0^{-1}\boldsymbol{\tau}_2 - \boldsymbol{\Psi}_0^{-1}\boldsymbol{\Psi}_1\boldsymbol{\Psi}_0^{-1}\boldsymbol{\tau}_1 + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^3), \tag{6.9}$$

where

$$\boldsymbol{\Psi}_0 = \boldsymbol{\Sigma}_1^2 \otimes \boldsymbol{I}_{m-r} - \boldsymbol{I}_r \otimes (\boldsymbol{\Sigma}_2^T \boldsymbol{\Sigma}_2),$$
$$\boldsymbol{\Psi}_1 = (\boldsymbol{\Sigma}_1 \boldsymbol{E}_{11} + \boldsymbol{E}_{11}^T \boldsymbol{\Sigma}_1) \otimes \boldsymbol{I}_{m-r} - \boldsymbol{I}_r \otimes (\boldsymbol{\Sigma}_2^T \boldsymbol{E}_{22} + \boldsymbol{E}_{22}^T \boldsymbol{\Sigma}_2),$$
$$\boldsymbol{\tau}_1 = -\operatorname{vec}(\boldsymbol{\Sigma}_2^T \boldsymbol{E}_{21} + \boldsymbol{E}_{12}^T \boldsymbol{\Sigma}_1), \qquad \boldsymbol{\tau}_2 = -\operatorname{vec}(\boldsymbol{E}_{22}^T \boldsymbol{E}_{21} + \boldsymbol{E}_{12}^T \boldsymbol{E}_{11}).$$

Corollary 6.1. *Suppose in Proposition 6.2,* $\boldsymbol{X}$ *has rank* r*, i.e.,* $\boldsymbol{\Sigma}_2 = \boldsymbol{0}$*. Then*

$$\boldsymbol{Q} = -\boldsymbol{E}_{21}\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{E}_{22}\boldsymbol{E}_{12}^T\boldsymbol{\Sigma}_1^{-2} + \boldsymbol{E}_{21}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{E}_{11}\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^3),$$
$$\boldsymbol{P} = \boldsymbol{E}_{12}^T\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{E}_{22}^T\boldsymbol{E}_{21}\boldsymbol{\Sigma}_1^{-2} - \boldsymbol{E}_{12}^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{E}_{11}^T\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^3).$$

Finally, we devote the rest of this section to discuss condition (6.5) in Proposition 6.2. As mentioned earlier, the singular subspaces corresponding to $\tilde{\boldsymbol{U}}_1$, $\tilde{\boldsymbol{U}}_2$, $\tilde{\boldsymbol{V}}_1$, and $\tilde{\boldsymbol{V}}_2$ are unique if and only if $\tilde{\sigma}_r > \tilde{\sigma}_{r+1}$. By Weyl's inequality (see Proposition 6.1), we have $|\tilde{\sigma}_{r+1} - \sigma_{r+1}| \leq \|\boldsymbol{\Delta}\|_2$. Since $\|\boldsymbol{\Delta}\|_2 < (\sigma_r - \sigma_{r+1})/2$ and $|\tilde{\sigma}_{r+1} - \sigma_{r+1}| \geq \tilde{\sigma}_{r+1} - \sigma_{r+1}$, one can further upper bound the $r+1$-th perturbed singular value by

$$\tilde{\sigma}_{r+1} < \sigma_{r+1} + \frac{\sigma_r - \sigma_{r+1}}{2} = \frac{\sigma_r + \sigma_{r+1}}{2}. \tag{6.10}$$

Following a similar argument, $|\tilde{\sigma}_r - \sigma_r| \leq \|\boldsymbol{\Delta}\|_2$ leads to

$$\tilde{\sigma}_r > \sigma_r - \frac{\sigma_r - \sigma_{r+1}}{2} = \frac{\sigma_r + \sigma_{r+1}}{2}. \tag{6.11}$$

It follows from (6.10) and (6.11) that the gap between $\tilde{\sigma}_r$ and $\tilde{\sigma}_{r+1}$ is strictly greater than 0:

$$\tilde{\sigma}_{r+1} < \frac{\sigma_r + \sigma_{r+1}}{2} < \tilde{\sigma}_r. \tag{6.12}$$

As mentioned in [82], condition (6.5) is more restrictive, but simpler, than the original condition specified in [190]. Based on the aforementioned preliminaries, we are ready to present our results.

6.4 Perturbation Expansions for the r-TSVD

This section presents perturbation expansion results for the r-TSVD operator. In order to guarantee the uniqueness of the expansions, we assume throughout the section that the r-th and $r+1$-th singular values are well-separated and the perturbation $\boldsymbol{\Delta}$ has small magnitude relative to $\boldsymbol{X}$.

Let us begin with a non-trivial result on the first-order perturbation expansion of the r-TSVD. The result is consistent with Theorem 25 from [64], in which Feppon and Lermusiaux utilized differential geometry to derive a closed-form expression for the directional derivative of the r-TSVD. Using tools from perturbation analysis, we are able to obtain the same result on the first-order perturbation expansion of $\mathcal{P}_r$. The additional benefit of the technique used here, as can be seen later, is that it can be leveraged to further derive the second-order perturbation expansion and the bound on the approximation error of the first-order expansion about a rank-r

matrix.

Theorem 6.1. *Assume $\sigma_r > \sigma_{r+1}$. Then, for some perturbation $\boldsymbol{\Delta} \in \mathbb{R}^{m\times n}$ such that $\|\boldsymbol{\Delta}\|_2 < \frac{\sigma_r - \sigma_{r+1}}{2}$, the first-order perturbation expansion of the r-TSVD about $\boldsymbol{X}$ is uniquely given by*[4]

$$\begin{aligned}\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) = \mathcal{P}_r(\boldsymbol{X}) + \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} \\ + \sum_{i=1}^{r}\sum_{j=r+1}^{n}\Bigg(\frac{\sigma_j^2}{\sigma_i^2-\sigma_j^2}(\boldsymbol{u}_i\boldsymbol{u}_i^T\boldsymbol{\Delta}\boldsymbol{v}_j\boldsymbol{v}_j^T + \boldsymbol{u}_j\boldsymbol{u}_j^T\boldsymbol{\Delta}\boldsymbol{v}_i\boldsymbol{v}_i^T) \\ + \frac{\sigma_i\sigma_j}{\sigma_i^2-\sigma_j^2}(\boldsymbol{u}_i\boldsymbol{v}_i^T\boldsymbol{\Delta}^T\boldsymbol{u}_j\boldsymbol{v}_j^T + \boldsymbol{u}_j\boldsymbol{v}_j^T\boldsymbol{\Delta}^T\boldsymbol{u}_i\boldsymbol{v}_i^T)\Bigg) + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2).\end{aligned} \tag{6.13}$$

The proof of Theorem 6.1 is based on perturbation expansions of the coefficient matrices $\boldsymbol{Q}$ and $\boldsymbol{P}$ in Proposition 6.3. Interested readers are encouraged to find out the details in Appendix 6.8.2. As mentioned earlier, the first-order term in (6.13) is equivalent to the directional derivative given by Theorem 25 in [64]:

$$\begin{aligned}&\boldsymbol{\nabla}_{\boldsymbol{\Delta}}\mathcal{P}_r(\boldsymbol{X}) = \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_1} + \boldsymbol{P}_{\boldsymbol{U}_1}\boldsymbol{\Delta} + \\ &\sum_{i=1}^{r}\sum_{j=r+1}^{m}\frac{\sigma_j}{\sigma_i^2-\sigma_j^2}\Big(\big(\sigma_i\boldsymbol{u}_j^T\boldsymbol{\Delta}\boldsymbol{v}_i + \sigma_j\boldsymbol{u}_i^T\boldsymbol{\Delta}\boldsymbol{v}_j\big)\boldsymbol{u}_j\boldsymbol{v}_i^T + \big(\sigma_j\boldsymbol{u}_j^T\boldsymbol{\Delta}\boldsymbol{v}_i + \sigma_i\boldsymbol{u}_i^T\boldsymbol{\Delta}\boldsymbol{v}_j\big)\boldsymbol{u}_i\boldsymbol{v}_j^T\Big).\end{aligned} \tag{6.14}$$

It is worthwhile to mention that we arrive at the first-order perturbation expansion in Theorem 6.1 while working independently on the error bounds for TSVD (see

[4]We recall that throughout this chapter we assume $m \geq n$.

Section 6.5).

Note that the condition $\|\boldsymbol{\Delta}\|_2 < (\sigma_r - \sigma_{r+1})/2$ guarantees a non-zero gap between the r-th and the $r+1$-th singular values of the perturbed matrix (see (6.12)), and hence guarantees $\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta})$ on the LHS of (6.13) is unique. At the same time, each term on the RHS of (6.13) is well-defined due to the uniqueness of singular subspaces associated with each group of singular values of $\boldsymbol{X}$. The term $\boldsymbol{\Delta} - \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2}$ can be viewed as the projection of $\boldsymbol{\Delta}$ onto the tangent space of the manifold of rank-r matrices [4]. On the other hand, the double summation stems from the curvature of this manifold when $\boldsymbol{X}$ does not lie on it (with rank greater than r). To demonstrate the first-order expansion in Theorem 6.1, let us consider the following examples.

Example 6.1. *Consider the matrix $\boldsymbol{X}$ with its SVD as follows:*

$$\boldsymbol{X} = \frac{1}{2}\begin{bmatrix} 4 & -4 & 7 \\ 0 & 0 & -9 \\ 4 & 8 & 1 \\ 8 & 4 & -1 \end{bmatrix} \tag{6.15}$$

$$= \left(\frac{1}{2}\begin{bmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}\right) \cdot \begin{bmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix} \cdot \left(\frac{1}{3}\begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ -2 & 2 & -1 \end{bmatrix}\right)^T . \tag{6.16}$$

In this example, note that $\sigma_1 = \sigma_2 > \sigma_3$. *From Definition 6.4, we have*

$$\mathcal{P}_2(\boldsymbol{X}) = \left(\frac{1}{2} \begin{bmatrix} -1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \right) \cdot \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix} \cdot \left(\frac{1}{3} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ -2 & 2 \end{bmatrix} \right)^T = \begin{bmatrix} 1 & -1 & 4 \\ -1 & 1 & -4 \\ 3 & 3 & 0 \\ 3 & 3 & 0 \end{bmatrix}. \tag{6.17}$$

In addition,

$$\boldsymbol{P}_{\boldsymbol{U}_2} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \qquad \boldsymbol{P}_{\boldsymbol{V}_2} = \frac{1}{9} \begin{bmatrix} 4 & -4 & -2 \\ -4 & 4 & 2 \\ -2 & 2 & 1 \end{bmatrix}. \tag{6.18}$$

For the perturbation

$$\boldsymbol{\Delta} = \frac{3}{200} \begin{bmatrix} 3 & 3 & -9 \\ -3 & -9 & 3 \\ 7 & 5 & -5 \\ -1 & 7 & -7 \end{bmatrix}, \ \textit{with} \ \|\boldsymbol{\Delta}\|_F = 0.2985 < \delta = 1.5, \tag{6.19}$$

(6.18) leads to

$$\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} = \frac{3}{200}\begin{bmatrix} 2 & -2 & -1 \\ 2 & -2 & -1 \\ 2 & -2 & -1 \\ -2 & 2 & 1 \end{bmatrix}. \tag{6.20}$$

Now the double summation in (6.13) can be represented as

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{\Delta}) = &\frac{1}{3}(\boldsymbol{u}_1\boldsymbol{u}_1^T\boldsymbol{\Delta}\boldsymbol{v}_3\boldsymbol{v}_3^T + \boldsymbol{u}_3\boldsymbol{u}_3^T\boldsymbol{\Delta}\boldsymbol{v}_1\boldsymbol{v}_1^T) + \frac{2}{3}(\boldsymbol{u}_1\boldsymbol{v}_1^T\boldsymbol{\Delta}^T\boldsymbol{u}_3\boldsymbol{v}_3^T + \boldsymbol{u}_3\boldsymbol{v}_3^T\boldsymbol{\Delta}^T\boldsymbol{u}_1\boldsymbol{v}_1^T) \\ &+ \frac{1}{3}(\boldsymbol{u}_2\boldsymbol{u}_2^T\boldsymbol{\Delta}\boldsymbol{v}_3\boldsymbol{v}_3^T + \boldsymbol{u}_3\boldsymbol{u}_3^T\boldsymbol{\Delta}\boldsymbol{v}_2\boldsymbol{v}_2^T) + \frac{2}{3}(\boldsymbol{u}_2\boldsymbol{v}_2^T\boldsymbol{\Delta}^T\boldsymbol{u}_3\boldsymbol{v}_3^T + \boldsymbol{u}_3\boldsymbol{v}_3^T\boldsymbol{\Delta}^T\boldsymbol{u}_2\boldsymbol{v}_2^T).\end{aligned}$$

While the singular vectors of $\boldsymbol{X}$ are not unique (due to $\sigma_1 = \sigma_2$), the singular subspaces of $\boldsymbol{X}$ are unique. Therefore, by representing $\boldsymbol{G}(\boldsymbol{\Delta})$ as

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{\Delta}) = &\frac{1}{3}(\boldsymbol{u}_1\boldsymbol{u}_1^T + \boldsymbol{u}_2\boldsymbol{u}_2^T)\boldsymbol{\Delta}\boldsymbol{v}_3\boldsymbol{v}_3^T + \frac{1}{3}\boldsymbol{u}_3\boldsymbol{u}_3^T\boldsymbol{\Delta}(\boldsymbol{v}_1\boldsymbol{v}_1^T + \boldsymbol{v}_2\boldsymbol{v}_2^T) \\ &+ \frac{2}{3}(\boldsymbol{u}_1\boldsymbol{v}_1^T + \boldsymbol{u}_2\boldsymbol{v}_2^T)\boldsymbol{\Delta}^T\boldsymbol{u}_3\boldsymbol{v}_3^T + \frac{2}{3}\boldsymbol{u}_3\boldsymbol{v}_3^T\boldsymbol{\Delta}^T(\boldsymbol{u}_1\boldsymbol{v}_1^T + \boldsymbol{u}_2\boldsymbol{v}_2^T),\end{aligned} \tag{6.21}$$

we observe that $\boldsymbol{G}(\boldsymbol{\Delta})$ is well-defined since $\boldsymbol{u}_1\boldsymbol{u}_1^T + \boldsymbol{u}_2\boldsymbol{u}_2^T$, $\boldsymbol{u}_3\boldsymbol{u}_3^T$, $\boldsymbol{v}_1\boldsymbol{v}_1^T + \boldsymbol{v}_2\boldsymbol{v}_2^T$,

$\boldsymbol{v}_3\boldsymbol{v}_3^T$, $\boldsymbol{u}_1\boldsymbol{v}_1^T + \boldsymbol{u}_2\boldsymbol{v}_2^T$, and $\boldsymbol{u}_3\boldsymbol{v}_3^T$ are all unique quantities, namely,

$$\boldsymbol{u}_1\boldsymbol{u}_1^T + \boldsymbol{u}_2\boldsymbol{u}_2^T = \boldsymbol{P}_{\boldsymbol{U}_1} = \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{u}_3\boldsymbol{u}_3^T = \frac{1}{4}\begin{bmatrix} 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 \end{bmatrix},$$

$$\boldsymbol{v}_1\boldsymbol{v}_1^T + \boldsymbol{v}_2\boldsymbol{v}_2^T = \boldsymbol{P}_{\boldsymbol{V}_1} = \frac{1}{9}\begin{bmatrix} 5 & 4 & 2 \\ 4 & 5 & -2 \\ 2 & -2 & 8 \end{bmatrix}, \qquad \boldsymbol{v}_3\boldsymbol{v}_3^T = \boldsymbol{P}_{\boldsymbol{V}_2} = \frac{1}{9}\begin{bmatrix} 4 & -4 & -2 \\ -4 & 4 & 2 \\ -2 & 2 & 1 \end{bmatrix},$$

$$\boldsymbol{u}_1\boldsymbol{v}_1^T + \boldsymbol{u}_2\boldsymbol{v}_2^T = \frac{1}{18}\begin{bmatrix} 3 & -3 & 12 \\ -3 & 3 & -12 \\ 9 & 9 & 0 \\ 9 & 9 & 0 \end{bmatrix}, \qquad \boldsymbol{u}_3\boldsymbol{v}_3^T = \frac{1}{6}\begin{bmatrix} 2 & -2 & -1 \\ 2 & -2 & -1 \\ -2 & 2 & 1 \\ 2 & -2 & -1 \end{bmatrix}. \quad (6.22)$$

Substituting the values of the 6 *aforementioned terms in (6.22) and the value of* $\boldsymbol{\Delta}$ *in (6.19) back into (6.21), we obtain*

$$\boldsymbol{G}(\boldsymbol{\Delta}) = \frac{1}{200}\begin{bmatrix} -6 & 3 & 0 \\ 2 & -5 & -4 \\ -2 & 5 & 4 \\ -6 & 3 & 0 \end{bmatrix}. \quad (6.23)$$

The substitution of (6.17), (6.19), (6.20), and (6.23) into (6.13) yields

$$\mathcal{P}_2(\boldsymbol{X}+\boldsymbol{\Delta}) = \begin{bmatrix} 0.9850 & -0.9100 & 3.8800 \\ -1.0650 & 0.8700 & -3.9600 \\ 3.0600 & 3.1300 & -0.0400 \\ 2.9850 & 3.0900 & -0.1200 \end{bmatrix} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \tag{6.24}$$

On the other hand, running a simple numerical evaluation by Definition 6.4, we can compute $\mathcal{P}_2(\boldsymbol{X}+\boldsymbol{\Delta})$ *and obtain*

$$\mathcal{P}_2(\boldsymbol{X}+\boldsymbol{\Delta}) = \begin{bmatrix} 0.9840 & -0.9088 & 3.8792 \\ -1.0632 & 0.8689 & -3.9615 \\ 3.0650 & 3.1284 & -0.0403 \\ 2.9870 & 3.0890 & -0.1213 \end{bmatrix}.$$

The approximation error of the first-order perturbation expansion has magnitude of 0.0043*, which is much smaller than the approximation error of the zero-order expansion, i.e.,* $\|\mathcal{P}_2(\boldsymbol{X}+\boldsymbol{\Delta}) - \mathcal{P}_2(\boldsymbol{X})\|_F = 0.3016$.

Example 6.2. *Let us consider a counter-example in which the condition* $\|\boldsymbol{\Delta}\|_2 <$

$(\sigma_r - \sigma_{r+1})/2$ is not satisfied. In particular, by setting

$$\boldsymbol{X} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \boldsymbol{\Delta} = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & -0.5 & 0 \\ 0 & 0 & 0.5 \\ 0 & 0 & 0 \end{bmatrix},$$

following similar calculation in Example 6.1 would yield

$$\mathcal{P}_2(\boldsymbol{X} + \boldsymbol{\Delta}) = \begin{bmatrix} 2.1 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2).$$

On the other hand, the 2-TSVD of $\boldsymbol{X} + \boldsymbol{\Delta}$ can either be

$$\mathcal{P}_2(\boldsymbol{X} + \boldsymbol{\Delta}) = \begin{bmatrix} 2.1 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \textit{or} \qquad \mathcal{P}_2(\boldsymbol{X} + \boldsymbol{\Delta}) = \begin{bmatrix} 2.1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1.5 \\ 0 & 0 & 0 \end{bmatrix}.$$

It can be seen that our first-order approximation is no longer accurate if the later truncation is considered.

One immediate consequence of Theorem 6.1 is when the matrix has exact rank r, the double summation on the RHS of (6.13) vanishes since $\sigma_j = 0$ for all $j > r$.

Thus, we obtain a simple expression for the first-order expansion of $\mathcal{P}_r(\cdot)$ about a rank-r matrix.

Corollary 6.2. *Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ be a rank-r matrix. Then, for some perturbation $\boldsymbol{\Delta} \in \mathbb{R}^{m\times n}$ such that $\|\boldsymbol{\Delta}\|_2 < \sigma_r/2$, the first-order perturbation expansion of the r-TSVD about $\boldsymbol{X}$ is uniquely given by*

$$\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) = \boldsymbol{X} + \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \tag{6.25}$$

We observe that while the first-order term depends on the perturbation $\boldsymbol{\Delta}$ and the two projections $\boldsymbol{P}_{\boldsymbol{U}_2}$ and $\boldsymbol{P}_{\boldsymbol{V}_2}$, it is independent of the singular values of $\boldsymbol{X}$. Motivated by the simple result in Corollary 6.2, we further study the second-order perturbation expansion of the r-TSVD about a rank-r matrix in the following theorem.

Theorem 6.2. *Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ be a rank-r matrix. Then, for some perturbation $\boldsymbol{\Delta} \in \mathbb{R}^{m\times n}$ such that $\|\boldsymbol{\Delta}\|_2 < \sigma_r/2$, the second-order perturbation expansion of the r-TSVD about $\boldsymbol{X}$ is uniquely given by*

$$\begin{aligned}\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) = \boldsymbol{X} + \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{X}^\dagger\boldsymbol{\Delta}^T\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger \\ + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^3).\end{aligned} \tag{6.26}$$

The proof of Theorem 6.2 is given in Appendix 6.8.3. The theorem states that $\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta})$ admits a simple second-order approximation that only depends on $\boldsymbol{P}_{\boldsymbol{U}_2}$, $\boldsymbol{P}_{\boldsymbol{V}_2}$, and $\boldsymbol{X}^\dagger$ in addition to $\boldsymbol{X}$ and $\boldsymbol{\Delta}$ themselves. Notice the dependence

of the three second-order terms on the RHS of (6.26) on the pseudo inverse of $\boldsymbol{X}$ indicates the first-order approximation is sensitive to the least singular value of $\boldsymbol{X}$. In the next section, we shall prove that the error bound for the first-order approximation of $\mathcal{P}_r(\boldsymbol{X} + \boldsymbol{\Delta})$ depends linearly on $1/\sigma_r$.

Remark 6.1. *The differentiability of $\mathcal{P}_r$ at a rank-r matrix, as shown in Corollary 2 and Theorem 2, matches with the well-known result in differential geometry that a projection onto the base of the normal bundle of any smooth manifold is a smooth map on the tubular neighborhood [123]. In particular, $\mathcal{P}_r$ is a classic smooth (C^∞) map in a small open neighborhood containing the manifold of rank-r matrices.*

Remark 6.2. *It is known that the r-TSVD is differentiable at any point (matrix) with a non-zero gap between the r-th and $r+1$-th singular values and hence, admits a first-order perturbation expansion about such point. While our result in Theorem 6.2 only considers a special case of rank-r matrices, we suspect there exists a second-order perturbation expansion of the r-TSVD about a matrix $\boldsymbol{X}$ with rank greater than r. However, given the complexity of the first-order expansion, it certainly requires more elaborate work. We leave this as a future research direction.*

6.5 Error Bounds for the r-TSVD

This section introduces upper bounds on the difference between the r-TSVD and its first-order approximation. While in Section 6.4 the perturbation expansions are derived under the assumption that $\|\boldsymbol{\Delta}\|_2 < (\sigma_r - \sigma_{r+1})/2$, the error bounds in this

section do not require this constraint and indeed they hold for $\boldsymbol{\Delta}$ with arbitrary magnitude. It is important to note that, without the constraint on the level of the perturbation, $\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta})$ may not be unique since there is no guarantee that $\tilde{\sigma}_r > \tilde{\sigma}_{r+1}$. The value of $\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta})$ in case $\tilde{\sigma}_r = \tilde{\sigma}_{r+1}$ depends on the choice of the singular subspace decomposition of $\tilde{\boldsymbol{X}} = \boldsymbol{X}+\boldsymbol{\Delta}$ (see Definition 6.2). Nevertheless, we shall provide error bounds that hold independent of the choice of decomposition.

Let us consider the first-order expansion in (6.25). One trivial bound on the approximation error can be derived as follows (see details in Appendix 6.8.4):

Lemma 6.1. *Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ be a rank-r matrix. For any $\boldsymbol{\Delta} \in \mathbb{R}^{m\times n}$ and any valid choice of subspace decomposition of $\boldsymbol{X}+\boldsymbol{\Delta}$, we have*

$$\|\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) - (\boldsymbol{X}+\boldsymbol{\Delta}-\boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2})\|_F \le \|\boldsymbol{X}\|_F + 2\|\boldsymbol{\Delta}\|_F .$$

Lemma 6.1 suggests that for large $\boldsymbol{\Delta}$, the approximation error grows at most linearly in the norm of $\boldsymbol{\Delta}$. However, for small $\boldsymbol{\Delta}$, the aforementioned bound is not tight since Corollary 6.2 implies the error should be in the order of $\|\boldsymbol{\Delta}\|_F^2$. In order to tighten the bound for the small perturbation, we need to develop a different approach that is more meticulous about intermediate inequalities. We state our main result regarding the global error bound on the first-order approximation of the r-TSVD as follows.

Theorem 6.3. *Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ be a rank-r matrix. Then, for any $\boldsymbol{\Delta} \in \mathbb{R}^{m\times n}$ and any valid choice of subspace decomposition of $\boldsymbol{X}+\boldsymbol{\Delta}$, the first-order Taylor*

expansion of the r-TSVD about $\boldsymbol{X}$ *is given by*

$$\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) = \boldsymbol{X} + \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}), \tag{6.27}$$

where there exists a universal constant $1 + 1/\sqrt{2} \leq c \leq 4(1+\sqrt{2})$ *such that*

$$\left\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\right\|_F \leq \frac{c}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F^2. \tag{6.28}$$

Furthermore, the following inequality holds

$$\left\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\right\|_F \leq 2(1+\sqrt{2})\left\|\boldsymbol{\Delta}\right\|_F \min\left\{\frac{2}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F, 1\right\}. \tag{6.29}$$

The proof of Theorem 6.3 is given in Appendix 6.8.5. It is noticeable that the first three terms on the RHS of (6.27) are uniquely given by the rank-r singular subspace decomposition of $\boldsymbol{X}$. On the contrary, the LHS may not be unique (e.g., when $\tilde{\sigma}_r = \tilde{\sigma}_{r+1}$) and hence, so does the residual $\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})$. However, it is interesting to note that the theorem makes no assumption on the norm of $\boldsymbol{\Delta}$, as well as the choice of the r-TSVD of $\boldsymbol{X}+\boldsymbol{\Delta}$. The bound on the residual (or the remainder) in Theorem 6.3 is similar to the Lagrange error bound in univariate first-order Taylor series. It not only asserts that the approximation error can grow no faster than a quadratic rate but also determines the constant attached to $\|\boldsymbol{\Delta}\|_F^2$. Furthermore, the bound depends only on the σ_r and $\|\boldsymbol{\Delta}\|_F$, as one may expect from the second-order perturbation expansion of the r-TSVD in Theorem 6.2.

Remark 6.3. *We conjecture but are unable to prove that the lower bound on*

c is tight, i.e., $c = 1 + 1/\sqrt{2}$. Partial result in this direction regarding $\boldsymbol{\Delta}$ of certain structure is also given in the proof of Theorem 6.3. In our numerical experiment, we ran multiple optimization procedures to maximize the quantity $\sigma_r \left\| \boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) \right\|_F / \left\| \boldsymbol{\Delta} \right\|_F^2$ with respect to $\boldsymbol{\Delta}$ and obtained the same constant $1+1/\sqrt{2}$.

The bound in (6.29) suggests an interesting behavior of the residual $\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})$. When the perturbation is small, the error depends quadratically on the magnitude of the perturbation and inversely proportional to the least singular value of $\boldsymbol{X}$. In particular, as σ_r approaches 0, the first-order approximation becomes less accurate. On the contrary, for large $\boldsymbol{\Delta}$, the upper bound is linear in the norm of $\boldsymbol{\Delta}$ and independent of σ_r. Compared to the bound in Lemma 6.1, we observe that the dependence on $\boldsymbol{X}$ is eliminated. Asymptotically as $\left\| \boldsymbol{\Delta} \right\|_F$ approaches ∞, the simple bound in the lemma becomes tighter than the bound in (6.29).

We conclude this section by describing the behavior of the residual term for small perturbations. While it is challenging to establish a tight bound on $\left\| \boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) \right\|_F$ (as a function of $\left\| \boldsymbol{\Delta} \right\|_F$) for large $\boldsymbol{\Delta}$, it is possible to project the first-order approximation error for small perturbation based on the knowledge of the second-order perturbation expansion of the r-TSVD (see Theorem 6.2). We provide the result in the following theorem, with the proof given in Appendix 6.8.6.

Theorem 6.4. *Asymptotically as $\left\| \boldsymbol{\Delta} \right\|_F$ approaches 0, the norm of the residual term in Theorem 6.3 can be upper-bounded tightly by*

$$\lim_{\epsilon \to 0^+} \sup_{\left\| \boldsymbol{\Delta} \right\|_F = \epsilon} \frac{\left\| \boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) \right\|_F}{\left\| \boldsymbol{\Delta} \right\|_F^2} = \frac{1}{\sigma_r \sqrt{3}}.$$

Remark 6.4. *While Theorem 6.1 provides the first-order perturbation expansion of the r-TSVD about an arbitrary matrix* $\boldsymbol{X}$ *with* $\sigma_r > \sigma_{r+1} \geq 0$*, extending Theorems 6.3 and 6.4 to that case remains to be one of our future research directions due to the difficulty of bounding the double summation in (6.13).*

6.6 An Application to Performance Analysis in Matrix Denoising

This section presents an application of our result to the performance analysis of the TSVD for matrix denoising. In many applications such as image denoising [85], multi-input multi-output (MIMO) channel estimation [134], collaborative filtering [118], low-rank procedures are often motivated by the following statistical model:

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta},$$

where $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is the unknown matrix with rank $r \leq \min(m, n)$ and $\boldsymbol{\Delta}$ is a random matrix whose entries are *i.i.d.* normally distributed with zero mean and σ^2-variance, i.e., $\Delta_{ij} \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. To denoise the data, the TSVD is applied to the noisy matrix $\tilde{\boldsymbol{X}}$ to obtain the following estimator:

$$\hat{\boldsymbol{X}} = \mathcal{P}_r(\tilde{\boldsymbol{X}}).$$

We would like to assess the mean squared error (MSE) of this estimator using our perturbation analysis of the TSVD. As a baseline for our analysis, we consider the

MSE of the noisy matrix $\tilde{\boldsymbol{X}}$:

$$\mathbb{E}\big[\Big\|\tilde{\boldsymbol{X}}-\boldsymbol{X}\Big\|_F^2\big]=\mathbb{E}\big[\|\boldsymbol{\Delta}\|_F^2\big]=\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}\big[\Delta_{ij}^2\big]=\sigma^2mn. \tag{6.30}$$

Next, we study the MSE of the estimator $\hat{\boldsymbol{X}}$, i.e., $\mathbb{E}[\Big\|\hat{\boldsymbol{X}}-\boldsymbol{X}\Big\|_F^2]$. To the best of our knowledge, there exists no closed-form expression of this quantity due to the non-linearity of the truncated singular value operator. In the following, we provide the first-order approximation, the second-order approximation, and the upper bound for $\mathbb{E}[\Big\|\hat{\boldsymbol{X}}-\boldsymbol{X}\Big\|_F^2]$ based on the results presented in this chapter.

1. **The first-order approximation:**

   Let $\hat{\boldsymbol{X}}_1=\boldsymbol{X}+\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}$ be the first-order approximation of $\hat{\boldsymbol{X}}$. We have

$$\begin{aligned}\mathbb{E}\Big[\Big\|\hat{\boldsymbol{X}}_1-\boldsymbol{X}\Big\|_F^2\Big]&=\mathbb{E}\Big[\|\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\|_F^2\Big]\\&=\mathbb{E}\Big[\|(\boldsymbol{I}_{mn}-\boldsymbol{P}_{\boldsymbol{V}_2}\otimes\boldsymbol{P}_{\boldsymbol{U}_2})\operatorname{vec}(\boldsymbol{\Delta})\|_2^2\Big]\\&\qquad\text{(by Lemma 6.8-2)}\\&=\mathbb{E}\Big[\big(\operatorname{vec}(\boldsymbol{\Delta})\big)^T(\boldsymbol{I}_{mn}-\boldsymbol{P}_{\boldsymbol{V}_2}\otimes\boldsymbol{P}_{\boldsymbol{U}_2})^T(\boldsymbol{I}_{mn}-\boldsymbol{P}_{\boldsymbol{V}_2}\otimes\boldsymbol{P}_{\boldsymbol{U}_2})\operatorname{vec}(\boldsymbol{\Delta})\Big].\end{aligned} \tag{6.31}$$

   Using the fact that $\boldsymbol{P}_{\boldsymbol{U}_2}$ and $\boldsymbol{P}_{\boldsymbol{V}_2}$ are projection matrices, $\boldsymbol{I}_{mn}-\boldsymbol{P}_{\boldsymbol{V}_2}\otimes\boldsymbol{P}_{\boldsymbol{U}_2}$ is also a projection matrix, and hence, $(\boldsymbol{I}_{mn}-\boldsymbol{P}_{\boldsymbol{V}_2}\otimes\boldsymbol{P}_{\boldsymbol{U}_2})^T(\boldsymbol{I}_{mn}-\boldsymbol{P}_{\boldsymbol{V}_2}\otimes\boldsymbol{P}_{\boldsymbol{U}_2})=$

$\boldsymbol{I}_{mn} - \boldsymbol{P}_{\boldsymbol{V}_2} \otimes \boldsymbol{P}_{\boldsymbol{U}_2}$. Thus, (6.31) can be further simplified as

$$\begin{aligned}
\mathbb{E}\Big[\Big\|\hat{\boldsymbol{X}}_1 - \boldsymbol{X}\Big\|_F^2\Big] &= \mathbb{E}\Big[\big(\operatorname{vec}(\boldsymbol{\Delta})\big)^T\big)(\boldsymbol{I}_{mn} - \boldsymbol{P}_{\boldsymbol{V}_2} \otimes \boldsymbol{P}_{\boldsymbol{U}_2})\operatorname{vec}(\boldsymbol{\Delta})\Big] \\
&= \mathbb{E}\Big[\operatorname{tr}\big((\boldsymbol{I}_{mn} - \boldsymbol{P}_{\boldsymbol{V}_2} \otimes \boldsymbol{P}_{\boldsymbol{U}_2})\operatorname{vec}(\boldsymbol{\Delta})\big(\operatorname{vec}(\boldsymbol{\Delta})\big)^T\big)\Big] \\
&\qquad \text{(by the cyclic property of the trace)} \\
&= \operatorname{tr}\Big((\boldsymbol{I}_{mn} - \boldsymbol{P}_{\boldsymbol{V}_2} \otimes \boldsymbol{P}_{\boldsymbol{U}_2})\mathbb{E}\big[\operatorname{vec}(\boldsymbol{\Delta})\big(\operatorname{vec}(\boldsymbol{\Delta})\big)^T\big]\Big).
\end{aligned}$$

Since $\Delta_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $\mathbb{E}\big[\operatorname{vec}(\boldsymbol{\Delta})\big(\operatorname{vec}(\boldsymbol{\Delta})\big)^T\big] = \sigma^2 \boldsymbol{I}_{mn}$. Thus,

$$\begin{aligned}
\mathbb{E}\Big[\Big\|\hat{\boldsymbol{X}}_1 - \boldsymbol{X}\Big\|_F^2\Big] &= \sigma^2 \operatorname{tr}(\boldsymbol{I}_{mn} - \boldsymbol{P}_{\boldsymbol{V}_2} \otimes \boldsymbol{P}_{\boldsymbol{U}_2}) \\
&= \sigma^2 \operatorname{tr}(\boldsymbol{I}_{mn}) - \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{V}_2}) \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{U}_2}) \\
&= \sigma^2 r(m + n - r), \qquad (6.32)
\end{aligned}$$

where the second equality uses Lemma 6.8-4 and the third equality stems from the fact that $\operatorname{tr}(\boldsymbol{P}_{\boldsymbol{U}_2}) = \operatorname{tr}(\boldsymbol{U}_2\boldsymbol{U}_2^T) = \operatorname{tr}(\boldsymbol{U}_2^T\boldsymbol{U}_2) = \operatorname{tr}(\boldsymbol{I}_{m-r}) = m - r$ (and similarly $\operatorname{tr}(\boldsymbol{P}_{\boldsymbol{V}_2}) = n - r$).

2. **The second-order approximation:**

   Let

$$\hat{\boldsymbol{X}}_2 = \boldsymbol{X} + \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{X}^\dagger\boldsymbol{\Delta}^T\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}$$

be the second-order approximation of $\hat{\boldsymbol{X}}$. We have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\hat{\boldsymbol{X}}_2 - \boldsymbol{X}\Big\|_F^2\Big] = \\
&\mathbb{E}\Big[\Big\|\boldsymbol{\Delta} - \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2} + \boldsymbol{X}^\dagger\boldsymbol{\Delta}^T\boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2} + \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2}\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger + \boldsymbol{P}_{U_2}\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{V_2}\Big\|_F^2\Big] \\
&= \mathbb{E}\big[\|\boldsymbol{\Delta} - \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2}\|_F^2\big] \\
&\qquad + \mathbb{E}\big[\Big\|\boldsymbol{X}^\dagger\boldsymbol{\Delta}^T\boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2} + \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2}\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger + \boldsymbol{P}_{U_2}\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{V_2}\Big\|_F^2\big] \\
&+ \mathbb{E}\Big[2\,\mathrm{tr}\big((\boldsymbol{\Delta} - \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2})^T(\boldsymbol{X}^\dagger\boldsymbol{\Delta}^T\boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2} \\
&\qquad + \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2}\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger + \boldsymbol{P}_{U_2}\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{V_2})\big)\Big]. \qquad (6.33)
\end{aligned}
$$

Since $\Delta_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, the expected value of the third-order term on the RHS of (6.33) is zero, i.e.,

$$
\begin{aligned}
&\mathbb{E}\Big[2\,\mathrm{tr}\big((\boldsymbol{\Delta} - \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2})^T(\boldsymbol{X}^\dagger\boldsymbol{\Delta}^T\boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2} \\
&\qquad\qquad + \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2}\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger + \boldsymbol{P}_{U_2}\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{V_2})\big)\Big] = 0.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\Big[\Big\|\hat{\boldsymbol{X}}_2 - \boldsymbol{X}\Big\|_F^2\Big] &= \mathbb{E}\big[\|\boldsymbol{\Delta} - \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2}\|_F^2\big] \\
&+ \mathbb{E}\big[\Big\|\boldsymbol{X}^\dagger\boldsymbol{\Delta}^T\boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2} + \boldsymbol{P}_{U_2}\boldsymbol{\Delta}\boldsymbol{P}_{V_2}\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger + \boldsymbol{P}_{U_2}\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{V_2}\Big\|_F^2\big].
\end{aligned}
$$

Since the first term on the RHS is given by (6.32), we proceed with the calcu-

lation of the second term on the RHS, i.e.,

$$\mathbb{E}\Big[\Big\|\boldsymbol{X}^{\dagger}\boldsymbol{\Delta}^{T}\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}+\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^{T}\boldsymbol{X}^{\dagger}+\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^{\dagger})^{T}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\Big\|_F^2\Big].$$

Since $\boldsymbol{X}^{\dagger}=\boldsymbol{P}_{\boldsymbol{U}_1}\boldsymbol{X}^{\dagger}\boldsymbol{P}_{\boldsymbol{V}_1}$, the three terms inside the norm are orthogonal to each other, i.e., their inner products are zero. Hence,

$$\begin{aligned}&\Big\|\boldsymbol{X}^{\dagger}\boldsymbol{\Delta}^{T}\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}+\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^{T}\boldsymbol{X}^{\dagger}+\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^{\dagger})^{T}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\Big\|_F^2\\&=\big\|\boldsymbol{X}^{\dagger}\boldsymbol{\Delta}^{T}\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\big\|_F^2+\big\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^{T}\boldsymbol{X}^{\dagger}\big\|_F^2+\Big\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^{\dagger})^{T}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\Big\|_F^2.\end{aligned}\tag{6.34}$$

Using the cyclic property of the trace and the idempotence property of $\boldsymbol{P}_{\boldsymbol{V}_2}$, the first term on the RHS of (6.34) can be computed as

$$\begin{aligned}\big\|\boldsymbol{X}^{\dagger}\boldsymbol{\Delta}^{T}\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\big\|_F^2&=\operatorname{tr}\big(\boldsymbol{X}^{\dagger}\boldsymbol{\Delta}^{T}\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^{T}\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^{\dagger})^{T}\big)\\&=\operatorname{tr}\big(\boldsymbol{\Delta}^{T}\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^{T}\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^{\dagger})^{T}\boldsymbol{X}^{\dagger}\big).\end{aligned}$$

Similarly, one can compute the second and the third terms on the RHS of (6.34),

then taking the expectation to obtain

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\boldsymbol{X}^\dagger\boldsymbol{\Delta}^T\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}+\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger+\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\Big\|_F^2\Big] \\
&\quad=\mathbb{E}\Big[\mathrm{tr}\big(\boldsymbol{\Delta}^T\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^T\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{X}^\dagger\big)\Big] \\
&\qquad\qquad+\mathbb{E}\Big[\mathrm{tr}\big(\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^T\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\big)\Big] \\
&\qquad\qquad+\mathbb{E}\Big[\mathrm{tr}\big(\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger\boldsymbol{\Delta}^T\boldsymbol{P}_{\boldsymbol{U}_2}\big)\Big]. \qquad (6.35)
\end{aligned}$$

Next, to compute the three terms on the RHS of (6.35), we consider the following lemma:

Lemma 6.2. *Assume the matrices $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$, and $\boldsymbol{D}$ in each of the following statements are of compatible dimensions such that the matrix product is valid. Then,*

(a) $\mathbb{E}\big[\mathrm{tr}(\boldsymbol{\Delta}^T\boldsymbol{A}\boldsymbol{\Delta}\boldsymbol{B}\boldsymbol{\Delta}^T\boldsymbol{C}\boldsymbol{\Delta}\boldsymbol{D})\big]=\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{C})\,\mathrm{tr}(\boldsymbol{B}^T\boldsymbol{D})+\mathrm{tr}(\boldsymbol{A}\boldsymbol{C})\,\mathrm{tr}(\boldsymbol{B})\,\mathrm{tr}(\boldsymbol{D})+\mathrm{tr}(\boldsymbol{B}\boldsymbol{D})\,\mathrm{tr}(\boldsymbol{A})\,\mathrm{tr}(\boldsymbol{C})),$

(b) $\mathbb{E}\big[\mathrm{tr}(\boldsymbol{\Delta}\boldsymbol{A}\boldsymbol{\Delta}\boldsymbol{B}\boldsymbol{\Delta}^T\boldsymbol{C}\boldsymbol{\Delta}^T\boldsymbol{D})\big]=\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{B}\boldsymbol{C}^T\boldsymbol{D})+\mathrm{tr}(\boldsymbol{D}\boldsymbol{C}\boldsymbol{B}\boldsymbol{A})+\mathrm{tr}(\boldsymbol{A}\boldsymbol{C})\,\mathrm{tr}(\boldsymbol{B})\,\mathrm{tr}(\boldsymbol{D}).$

The proof of Lemma 6.2 follows a similar derivation of the fourth-moment properties in [161] and hence is omitted. Applying Lemma 6.2 to the RHS of (6.35)

and using the orthogonality between $\boldsymbol{X}^\dagger$ and $\boldsymbol{P}_{\boldsymbol{U}_2}, \boldsymbol{P}_{\boldsymbol{V}_2}$, we obtain

$$\begin{aligned}
&\mathbb{E}\big[\, \Big\| \boldsymbol{X}^\dagger \boldsymbol{\Delta}^T \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} \boldsymbol{\Delta}^T \boldsymbol{X}^\dagger + \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} (\boldsymbol{X}^\dagger)^T \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} \Big\|_F^2 \big] \\
&\quad = \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{U}_2}) \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{V}_2}) \operatorname{tr}((\boldsymbol{X}^\dagger)^T \boldsymbol{X}^\dagger) + \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{V}_2}) \operatorname{tr}(\boldsymbol{X}^\dagger (\boldsymbol{X}^\dagger)^T) \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{U}_2}) \\
&\qquad\qquad + \operatorname{tr}((\boldsymbol{X}^\dagger)^T \boldsymbol{X}^\dagger) \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{V}_2}) \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{U}_2}) \\
&\quad = 3\sigma^4 (m-r)(n-r) \left\| \boldsymbol{X}^\dagger \right\|_F^2 , \qquad (6.36)
\end{aligned}$$

where the last equality stems from $\operatorname{tr}(\boldsymbol{P}_{\boldsymbol{U}_2}) = m - r$ and $\operatorname{tr}(\boldsymbol{P}_{\boldsymbol{V}_2}) = n - r$. Substituting (6.32) and (6.36) into (6.33) yields

$$\mathbb{E}\Big[\Big\| \hat{\boldsymbol{X}}_2 - \boldsymbol{X} \Big\|_F^2 \Big] = \sigma^2 r(m+n-r) + 3\sigma^4 (m-r)(n-r) \left\| \boldsymbol{X}^\dagger \right\|_F^2 . \qquad (6.37)$$

3. **The upper bound:**

From Corollary 6.2, we have

$$\hat{\boldsymbol{X}} - \boldsymbol{X} = \mathcal{P}_r(\tilde{\boldsymbol{X}}) - \boldsymbol{X} = \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}).$$

Hence, by the triangle inequality, it holds that

$$\Big\| \hat{\boldsymbol{X}} - \boldsymbol{X} \Big\|_F \leq \| \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} \|_F + \| \boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) \|_F .$$

Taking the expectation of the squared norm yields

$$\mathbb{E}\Big[\Big\|\hat{\boldsymbol{X}}-\boldsymbol{X}\Big\|_F^2\Big] \leq \mathbb{E}\Big[\Big(\|\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\|_F + \|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F\Big)^2\Big]. \tag{6.38}$$

Applying Minkowski inequality [93], we can bound the RHS of (6.36) as

$$\begin{aligned}\mathbb{E}\Big[\Big(\|\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\|_F + \|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F\Big)^2\Big] \\ \leq \left(\sqrt{\mathbb{E}\Big[\|\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\|_F^2\Big]} + \sqrt{\mathbb{E}\Big[\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F^2\Big]}\right)^2.\end{aligned} \tag{6.39}$$

From (6.29), we can bound $\mathbb{E}\Big[\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F^2\Big]$ by

$$\begin{aligned}\mathbb{E}\Big[\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F^2\Big] &\leq \mathbb{E}\Big[\big(2(1+\sqrt{2})\min\big\{\frac{2}{\sigma_r}\|\boldsymbol{\Delta}\|_F^2, \|\boldsymbol{\Delta}\|_F\big\}\big)^2\Big] \\ &\leq \min\left\{\Big(4(1+\sqrt{2})\frac{\sigma}{\sigma_r}\Big)^2\mathbb{E}\Big[\|\boldsymbol{\Delta}\|_F^4\Big], \Big(2(1+\sqrt{2})\Big)^2\mathbb{E}\Big[\|\boldsymbol{\Delta}\|_F^2\Big]\right\},\end{aligned} \tag{6.40}$$

where the last inequality is a special case of Jensen's inequality [93] with the minimum of two linear functions as a concave function. The fourth-order term on the RHS of (6.40) can be computed as

$$\begin{aligned}\mathbb{E}\Big[\|\boldsymbol{\Delta}\|_F^4\Big] &= \mathbb{E}\Big[\big(\sum_{i=1}^{m}\sum_{j=1}^{n}\Delta_{ij}^2\big)^2\Big] = \sum_{i,j,k,l}\mathbb{E}\Big[\Delta_{ij}^2\Delta_{kl}^2\Big] = \sigma^4\sum_{i,j,k,l}(1+2\delta_{ik}\delta_{jl}) \\ &= \sigma^4(m^2n^2+2mn).\end{aligned} \tag{6.41}$$

Substituting (6.32) and (6.41) back into (6.40), then taking the square root, we have

$$\sqrt{\mathbb{E}\Big[\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F^2\Big]} \leq \min\Bigg\{4(1+\sqrt{2})\frac{\sigma}{\sigma_r}\sqrt{m^2n^2+2mn}, 2(1+\sqrt{2})\sqrt{mn}\Bigg\}.$$

Substituting the bound in the last inequality and the equality in (6.32) into (6.39), we obtain the upper bound as

$$\begin{aligned}\mathbb{E}\Big[\Big\|\hat{\boldsymbol{X}}-\boldsymbol{X}\Big\|_F^2\Big] \leq \sigma^2\Bigg(&\sqrt{r(m+n-r)} \\ &+\min\Bigg\{4(1+\sqrt{2})\frac{\sigma}{\sigma_r}\sqrt{m^2n^2+2mn}, 2(1+\sqrt{2})\sqrt{mn}\Bigg\}\Bigg)^2. \end{aligned} \tag{6.42}$$

Due to the nature of the bound given in (6.29), the bound in (6.42) is taken as the minimum between a component that is linear in the norm of $\boldsymbol{\Delta}$ and a component that is quadratic in the norm of $\boldsymbol{\Delta}$.

Remark 6.5. *Asymptotically as $\sigma \to 0$, all the ratios of the first-order approximation (6.32), the second-order approximation (6.37), and the upper bound (6.42) to the MSE of the noisy matrix (6.30) converge to $r(m+n-r)/mn$. In general, this ratio is less than or equal to 1, however, in low-rank scenarios it can be significantly smaller. This indicates the TSVD estimator is effective in noise reduction when the noise is small, especially when the matrix $\boldsymbol{X}$ has low rank.*

Remark 6.6. *The upper bound in (6.42) attains the same value of the baseline*

$\sigma^2 mn$ *when* $\sigma = \sigma_2$*, where*

$$\sigma_2 = \frac{\sigma_r\big(\sqrt{mn} - \sqrt{r(m+n-r)}\big)}{4(1+\sqrt{2})\sqrt{m^2n^2+2mn}}, \tag{6.43}$$

guaranteeing the superiority of the upper bound over the baseline in the case $\sigma < \sigma_2$*.*

Remark 6.7. *Let us define the* ρ*-knee point between two increasing functions of* σ*, e.g.,* $f(\sigma)$ *and* $g(\sigma)$*, as the point at which* $f(\sigma) = \rho g(\sigma)$ *(for* $\rho > 1$*). Then, the* ρ*-knee point between the upper bound (6.42) and the first-order approximation (6.32) can be determined by*

$$\sigma_1 = \frac{\sigma_r(\sqrt{\rho}-1)\sqrt{r(m+n-r)}}{4(1+\sqrt{2})\sqrt{m^2n^2+2mn}}. \tag{6.44}$$

In addition, the ρ*-knee point between the second-order approximation (6.37) and the first-order approximation (6.32) is given by*

$$\sigma_3 = \sqrt{\frac{(\rho-1)r(m+n-r)}{3(m-r)(n-r)\left\|\boldsymbol{X}^\dagger\right\|_F^2}} > \sigma_1. \tag{6.45}$$

Figure 6.1 demonstrates the aforementioned analysis on the performance of the TSVD-based estimator for matrix denoising through a numerical experiment. **Data generation.** We generate a matrix $\boldsymbol{X}$ with $m = 100, n = 80$, and $r = 3$ by (i) taking the product of two random matrices, whose entries are *i.i.d.* normally distributed $\mathcal{N}(0,1)$, of sizes 100×3 and 3×80, respectively; (ii) and dividing each entry of the obtained matrix by its Frobenius norm such that the resulting matrix satisfies $\|\boldsymbol{X}\|_F = 1$. In the experiment, we consider 51 values of σ in the interval

of $[10^{-6}, 10^{0}]$, namely $\sigma \in \{10^{-6}, 10^{-5.88}, 10^{-5.76}, \ldots, 10^{0}\}$. For each value of σ, we compute the following quantities:

1. the empirical MSE of the TSVD-based estimator $\mathbb{E}[\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}\right\|_F^2]$ by averaging the quantity $\|\mathcal{P}_r(\boldsymbol{X} + \boldsymbol{\Delta}) - \boldsymbol{X}\|_F^2$ over 1000 *i.i.d.* instances of $\boldsymbol{\Delta}$,
2. the MSE of the noisy matrix given in (6.30),
3. the first-order approximation of the MSE of the TSVD-based estimator given in (6.32),
4. the second-order approximation of the MSE of the TSVD-based estimator given in (6.33),
5. the upper bound on the MSE of the TSVD-based estimator given in (6.42).

We display each of the aforementioned quantities as a function of σ in Fig. 6.1. In addition, we calculate the points corresponding to σ_1, σ_2 and σ_3 using (6.44), (6.43), and (6.45), respectively, with $\rho = 1.1$, and include them in Fig. 6.1. **Results and Analysis.** It can be observed from the plot that the empirical MSE of the TSVD-based estimator (solid blue) increases quadratically as a function of σ (in the log-log scale, it appears as a straight line with slope equal to 2). The first-order approximation (dash-dotted yellow) and the second-order approximation (dash-dotted purple) match the empirical average well for $\sigma < \sigma_3 \approx 10^{-2}$. In this range of σ, all of the three aforementioned quantities are lower than the MSE of the noisy matrix (solid red). On the other hand, the upper bound (solid green) holds tightly when $\sigma < \sigma_1 \approx 10^{-4}$, providing an efficient guarantee on the performance

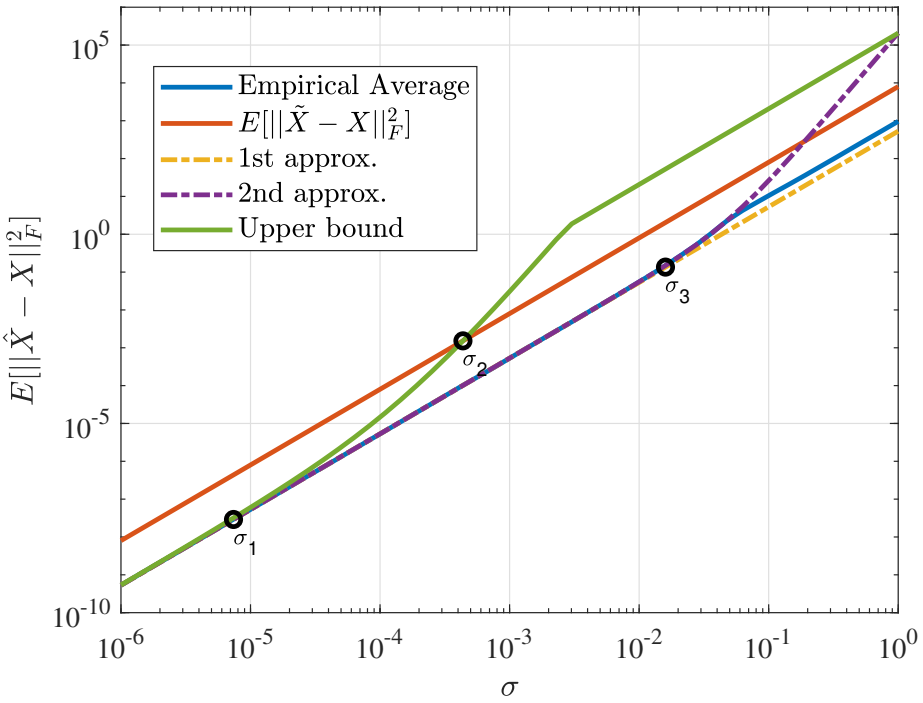


Figure 6.1: The MSE of the TSVD-based estimator $\hat{\boldsymbol{X}}$ for matrix denoising as a function of σ. The solid blue line represents the empirical estimate of MSE of $\hat{\boldsymbol{X}}$, i.e., $\mathbb{E}[\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}\right\|_F^2]$. The solid red line is the MSE of the noisy matrix, i.e., $\mathbb{E}[\left\|\tilde{\boldsymbol{X}} - \boldsymbol{X}\right\|_F^2] = \sigma^2 mn$. The dash-dotted yellow line and the dash-dotted purple line represent the first-order and second-order approximations of $\mathbb{E}[\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}\right\|_F^2]$, i.e., $\mathbb{E}[\left\|\hat{\boldsymbol{X}}_1 - \boldsymbol{X}\right\|_F^2]$ and $\mathbb{E}[\left\|\hat{\boldsymbol{X}}_2 - \boldsymbol{X}\right\|_F^2]$, respectively. The solid green line is the upper bound on $\mathbb{E}[\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}\right\|_F^2]$ given in (6.42). The knee-points σ_1 and σ_3 represent the value of σ for which the upper-bound and the second-order approximation deviate from the first order approximation by more than 10%, obtained by (6.44) and (6.45) with $\rho = 1.1$. The point σ_2 is the intersection between the upper bound and the MSE of the noisy matrix, given by (6.43).

of the TSVD-based estimator for denoising with the presence of small additive noises. However, as the noise variance increases, the upper bound appears loose, exceeding the MSE of the noisy matrix when $\sigma > \sigma_2 \approx 4 \times 10^{-4}$. The bound is developed for the worst-case noise scenario, in which the noise is adversarially selected to yield the largest perturbation error (see the proof of Theorem 6.3 in Appendix 6.8.5) and not for the random noise case. Consequently, it is far more conservative, predicting a larger MSE than the actual MSE of the TSVD-based estimator. Developing bounds for average-case scenario is a potential direction for future research.

6.7 Conclusion

In this chapter, we derived a first-order perturbation expansion for the singular value truncation. When the underlying matrix has exact rank-r, we showed that the first-order approximation can be greatly simplified and further introduced a simple expression of the second-order perturbation expansion for the r-TSVD. Next, we proposed an error bound on the first-order approximation of the r-TSVD about a rank-r matrix. Our bound is universal in the sense that it holds for perturbation matrices with an arbitrary norm. Two open questions raised by our analysis are: (i) when the underlying matrix has arbitrary rank, whether there exists an explicit expression for the second-order perturbation expansion of the TSVD; (ii) and given the result in Theorem 6.1, whether it is possible to establish a global error bound on the first-order approximation of the r-TSVD.

6.8 Appendix

6.8.1 Auxiliary Lemmas

This section summarizes some trivial results that will be used regularly in our subsequent derivation. The proofs of Lemmas 6.3-6.7 can be found in [151] - Chapter 5. The proof of Lemma 6.8 can be found in [81] - Chapter 2.

Lemma 6.3. *Assume the same setting as in Definition 6.2. The following statements hold:*

1. $\boldsymbol{U}_1^T\boldsymbol{U}_1 = \boldsymbol{V}_1^T\boldsymbol{V}_1 = \boldsymbol{I}_r$, $\boldsymbol{U}_2^T\boldsymbol{U}_2 = \boldsymbol{I}_{m-r}$, *and* $\boldsymbol{V}_2^T\boldsymbol{V}_2 = \boldsymbol{I}_{n-r}$,
2. $\boldsymbol{U}_1^T\boldsymbol{U}_2 = \boldsymbol{0}_{r\times(m-r)}$ *and* $\boldsymbol{V}_1^T\boldsymbol{V}_2 = \boldsymbol{0}_{r\times(n-r)}$,
3. $\boldsymbol{P}_{\boldsymbol{U}_1}\boldsymbol{P}_{\boldsymbol{U}_2} = \boldsymbol{0}$ *and* $\boldsymbol{P}_{\boldsymbol{V}_1}\boldsymbol{P}_{\boldsymbol{V}_2} = \boldsymbol{0}$.

Furthermore, if $\boldsymbol{X}$ *has rank* r*, then*

1. $\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{X} = \boldsymbol{0}$ *and* $\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{V}_2} = \boldsymbol{0}$,
2. $\boldsymbol{X} = \boldsymbol{P}_{\boldsymbol{U}_1}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{P}_{\boldsymbol{V}_1}$,
3. $\boldsymbol{X}(\boldsymbol{X}^\dagger)^T = \boldsymbol{P}_{\boldsymbol{U}_1}$ *and* $\boldsymbol{X}^T\boldsymbol{X}^\dagger = \boldsymbol{P}_{\boldsymbol{V}_1}$.

Lemma 6.4. *Assume the same setting as in Definition 6.2. The following statements hold:*

1. $\mathcal{P}_r(\boldsymbol{X}) = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T = \boldsymbol{P}_{\boldsymbol{U}_1}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{P}_{\boldsymbol{V}_1} = \boldsymbol{P}_{\boldsymbol{U}_1}\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{V}_1}$,
2. $\boldsymbol{X} - \mathcal{P}_r(\boldsymbol{X}) = \boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T = \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{P}_{\boldsymbol{V}_2} = \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{V}_2}$.

Lemma 6.5. *For any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with compatible dimensions, the following inequalities hold*

$$\|\boldsymbol{AB}\|_2 \leq \|\boldsymbol{AB}\|_F \leq \min\{\|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_2, \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_F\} \leq \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F .$$

Lemma 6.6. *(Pythagoras theorem for Frobenius norm) For any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ such that $\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{B}) = 0$, it holds that*

$$\|\boldsymbol{A} + \boldsymbol{B}\|_F = \sqrt{\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2}.$$

The matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in this case are said to be orthogonal to each other.

Lemma 6.7. *Let $\boldsymbol{U}$ be a semi-orthogonal matrix with orthonormal columns and $\boldsymbol{P_U} = \boldsymbol{U}\boldsymbol{U}^T$. Then, for any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ that have compatible dimensions with $\boldsymbol{U}$, the followings hold*

1. $\|\boldsymbol{UA}\|_2 = \|\boldsymbol{A}\|_2$ *and* $\|\boldsymbol{UA}\|_F = \|\boldsymbol{A}\|_F$,
2. $\|\boldsymbol{BU}\|_2 = \|\boldsymbol{BP_U}\|_2 \leq \|\boldsymbol{B}\|_2$ *and* $\|\boldsymbol{BU}\|_F = \|\boldsymbol{BP_U}\|_F \leq \|\boldsymbol{B}\|_F$.

Lemma 6.8. *For any matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ with compatible dimensions such that the matrix products are valid, the following holds*

1. $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{AC}) \otimes (\boldsymbol{BD})$,
2. $\operatorname{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^T \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{B})$,
3. $\|\boldsymbol{A} \otimes \boldsymbol{B}\|_F = \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F$,
4. $\operatorname{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) \operatorname{tr}(\boldsymbol{B})$.

6.8.2 Proof of Theorem 6.1

Recall that in this proof, we consider a matrix $\boldsymbol{X}$ having rank greater than or equal to r. With a slight abuse of notation, let us define $\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})$ as follows:

$$\begin{aligned}\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) &= \mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) - \big(\mathcal{P}_r(\boldsymbol{X}) + \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\big) &(6.46)\\ &= \big(\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) - (\boldsymbol{X}+\boldsymbol{\Delta})\big) + \big(\boldsymbol{X} - \mathcal{P}_r(\boldsymbol{X})\big) + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}. &(6.47)\end{aligned}$$

Since $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta}$, applying Lemma 6.4 to (6.47) yields

$$\begin{aligned}\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) &= -\boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2} + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\\ &= -\boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2} + \boldsymbol{P}_{\boldsymbol{U}_2}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{V}_2}.\end{aligned}$$

Denote $\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}} = \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2} - \boldsymbol{P}_{\boldsymbol{U}_2}$ and $\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} = \boldsymbol{P}_{\tilde{\boldsymbol{V}}_2} - \boldsymbol{P}_{\boldsymbol{U}_2}$. By rewriting $\boldsymbol{P}_{\tilde{\boldsymbol{U}}_2} = \boldsymbol{P}_{\boldsymbol{U}_2} + \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}$ and $\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2} = \boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}$, we can further simplify the last equation as

$$\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) = -\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{V}_2} - \boldsymbol{P}_{\boldsymbol{U}_2}\tilde{\boldsymbol{X}}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} - \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\tilde{\boldsymbol{X}}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}. \tag{6.48}$$

Lemma 6.9. *The perturbations of singular subspaces satisfy*

$$\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}} = \boldsymbol{U}_1\boldsymbol{Q}^T\boldsymbol{U}_2^T + \boldsymbol{U}_2\boldsymbol{Q}\boldsymbol{U}_1^T + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2), \tag{6.49a}$$

$$\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} = -\boldsymbol{V}_1\boldsymbol{P}^T\boldsymbol{V}_2^T - \boldsymbol{V}_2\boldsymbol{P}\boldsymbol{V}_1^T + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \tag{6.49b}$$

The proof of Lemma 6.9 is given at the end of this section. From this lemma, it is

clear that $\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}$ and $\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}}$ are in the order of $\|\boldsymbol{\Delta}\|_F$. Substituting $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta}$ into (6.48) and collecting second-order terms yield

$$\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) = -\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}} \boldsymbol{X} \boldsymbol{P}_{V_2} - \boldsymbol{P}_{U_2} \boldsymbol{X} \boldsymbol{\delta}_{\boldsymbol{P}_{V_2}} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \tag{6.50}$$

Substituting (6.49a) into the first term on the RHS of (6.50), we obtain

$$\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}} \boldsymbol{X} \boldsymbol{P}_{V_2} = \big(\boldsymbol{U}_1 \boldsymbol{Q}^T \boldsymbol{U}_2^T + \boldsymbol{U}_2 \boldsymbol{Q} \boldsymbol{U}_1^T\big) \boldsymbol{U}_2 \boldsymbol{\Sigma}_2 \boldsymbol{V}_2^T + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2).$$

Since $\boldsymbol{U}_2^T \boldsymbol{U}_2 = \boldsymbol{I}$ and $\boldsymbol{U}_1^T \boldsymbol{U}_2 = \boldsymbol{0}$, we further have

$$\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}} \boldsymbol{X} \boldsymbol{P}_{V_2} = \boldsymbol{U}_1 \boldsymbol{Q}^T \boldsymbol{\Sigma}_2 \boldsymbol{V}_2^T + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \tag{6.51}$$

Similarly, the second term on the RHS of (6.50) can be represented as

$$\boldsymbol{P}_{U_2} \boldsymbol{X} \boldsymbol{\delta}_{\boldsymbol{P}_{V_2}} = -\boldsymbol{U}_2 \boldsymbol{\Sigma}_2 \boldsymbol{P} \boldsymbol{V}_1^T + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \tag{6.52}$$

Substituting (6.51) and (6.52) back into (6.50), we have

$$\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) = -\boldsymbol{U}_1 \boldsymbol{Q}^T \boldsymbol{\Sigma}_2 \boldsymbol{V}_2^T + \boldsymbol{U}_2 \boldsymbol{\Sigma}_2 \boldsymbol{P} \boldsymbol{V}_1^T + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \tag{6.53}$$

Now we can vectorize (6.53) and apply Lemma 6.8 to obtain

$$\operatorname{vec}\big(\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\big) = (\boldsymbol{V}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{U}_1)\operatorname{vec}(-\boldsymbol{Q}^T) + (\boldsymbol{V}_1 \otimes \boldsymbol{U}_2\boldsymbol{\Sigma}_2)\operatorname{vec}(\boldsymbol{P}) + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \tag{6.54}$$

Let us now consider each term on the RHS of (6.54). From Proposition 6.3, it follows that

$$\operatorname{vec}(-\boldsymbol{Q}^T) = (\boldsymbol{I}_{m-r} \otimes \boldsymbol{\Sigma}_1^2 - \boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{I}_r)^{-1}\operatorname{vec}(\boldsymbol{E}_{12}\boldsymbol{\Sigma}_2^T + \boldsymbol{\Sigma}_1\boldsymbol{E}_{21}^T) + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \tag{6.55}$$

Replacing $\boldsymbol{E}_{ij} = \boldsymbol{U}_i^T\boldsymbol{\Delta}\boldsymbol{V}_j$, for $i, j \in \{1, 2\}$, and using Lemma 6.8, (6.55) becomes

$$\begin{aligned}\operatorname{vec}(-\boldsymbol{Q}^T) = (\boldsymbol{I}_{m-r} \otimes \boldsymbol{\Sigma}_1^2 - \boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{I}_r)^{-1}\big((\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T \otimes \boldsymbol{U}_1^T)\operatorname{vec}(\boldsymbol{\Delta}) \\ + (\boldsymbol{U}_2^T \otimes \boldsymbol{\Sigma}_1\boldsymbol{V}_1^T)\operatorname{vec}(\boldsymbol{\Delta}^T)\big) + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2).\end{aligned} \tag{6.56}$$

Since $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are diagonal, so is $(\boldsymbol{I}_{m-r} \otimes \boldsymbol{\Sigma}_1^2 - \boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{I}_r)^{-1}$. The following lemma provides an insight into the structure of this inversion.

Lemma 6.10. *Let* $\boldsymbol{D} = (\boldsymbol{I}_{m-r} \otimes \boldsymbol{\Sigma}_1^2 - \boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{I}_r)^{-1}$. *Then*

$$\boldsymbol{D} = \sum_{i=1}^{r}\sum_{k=1}^{m-r} d_{ik}\big(\boldsymbol{e}_k^{m-r}(\boldsymbol{e}_k^{m-r})^T\big) \otimes \big(\boldsymbol{e}_i^r(\boldsymbol{e}_i^r)^T\big),$$

where $d_{ik} = \frac{1}{\sigma_i^2 - \sigma_{r+k}^2}$, *for* $i = 1, \ldots, r$ *and* $k = 1, \ldots, m-r$.

The proof of Lemma 6.10 is given at the end of this section. Now using Lemma 6.10

and left-multiplying both sides of (6.56) by $(\boldsymbol{V}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{U}_1)$, we obtain

$$
\begin{aligned}
&(\boldsymbol{V}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{U}_1)\,\mathrm{vec}(-\boldsymbol{Q}^T) \\
&= \sum_{i=1}^{r}\sum_{k=1}^{m-r} d_{ik}(\boldsymbol{V}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{U}_1)\Big(\big(\boldsymbol{e}_k^{m-r}(\boldsymbol{e}_k^{m-r})^T\big) \otimes \big(\boldsymbol{e}_i^{r}(\boldsymbol{e}_i^{r})^T\big)\Big)(\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T \otimes \boldsymbol{U}_1^T)\,\mathrm{vec}(\boldsymbol{\Delta}) \\
&+ \sum_{i=1}^{r}\sum_{k=1}^{m-r} d_{ik}(\boldsymbol{V}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{U}_1)\Big(\big(\boldsymbol{e}_k^{m-r}(\boldsymbol{e}_k^{m-r})^T\big) \otimes \big(\boldsymbol{e}_i^{r}(\boldsymbol{e}_i^{r})^T\big)\Big)(\boldsymbol{U}_2^T \otimes \boldsymbol{\Sigma}_1\boldsymbol{V}_1^T)\,\mathrm{vec}(\boldsymbol{\Delta}^T) \\
&+ \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \qquad (6.57)
\end{aligned}
$$

Moreover, applying Lemma 6.8-1, we have

$$
\begin{aligned}
(\boldsymbol{V}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{U}_1)\Big(\big(\boldsymbol{e}_k^{m-r}(\boldsymbol{e}_k^{m-r})^T\big) \otimes \big(\boldsymbol{e}_i^{r}(\boldsymbol{e}_i^{r})^T\big)\Big)(\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T \otimes \boldsymbol{U}_1^T) \\
= \big(\boldsymbol{V}_2\boldsymbol{\Sigma}_2^T\boldsymbol{e}_k^{m-r}(\boldsymbol{e}_k^{m-r})^T\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T\big) \otimes \big(\boldsymbol{U}_1\boldsymbol{e}_i^{r}(\boldsymbol{e}_i^{r})^T\boldsymbol{U}_1^T\big) \\
= \sigma_{r+k}^2(\boldsymbol{v}_{r+k}\boldsymbol{v}_{r+k}^T) \otimes (\boldsymbol{u}_i\boldsymbol{u}_i^T), \qquad (6.58)
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
(\boldsymbol{V}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{U}_1)\Big(\big(\boldsymbol{e}_k^{m-r}(\boldsymbol{e}_k^{m-r})^T\big) \otimes \big(\boldsymbol{e}_i^{r}(\boldsymbol{e}_i^{r})^T\big)\Big)(\boldsymbol{U}_2^T \otimes \boldsymbol{\Sigma}_1\boldsymbol{V}_1^T) \\
= \sigma_i\sigma_{r+k}(\boldsymbol{v}_{r+k}\boldsymbol{u}_{r+k}^T) \otimes (\boldsymbol{u}_i\boldsymbol{v}_i^T). \qquad (6.59)
\end{aligned}
$$

Substituting (6.58) and (6.59) back into (6.57) and performing a change of variable

$j = r + k$, we obtain

$$(\boldsymbol{V}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{U}_1)\,\mathrm{vec}(-\boldsymbol{Q}^T) = \sum_{i=1}^{r}\sum_{j=r+1}^{m} \frac{\sigma_j^2}{\sigma_i^2 - \sigma_j^2}(\boldsymbol{v}_j\boldsymbol{v}_j^T) \otimes (\boldsymbol{u}_i\boldsymbol{u}_i^T)\,\mathrm{vec}(\boldsymbol{\Delta}) + \sum_{i=1}^{r}\sum_{j=r+1}^{m} \frac{\sigma_i\sigma_j}{\sigma_i^2 - \sigma_j^2}(\boldsymbol{v}_j\boldsymbol{u}_j^T) \otimes (\boldsymbol{u}_i\boldsymbol{v}_i^T)\,\mathrm{vec}(\boldsymbol{\Delta}^T) + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \quad (6.60)$$

Following a similar derivation, we also have

$$(\boldsymbol{V}_1 \otimes \boldsymbol{U}_2\boldsymbol{\Sigma}_2)\,\mathrm{vec}(\boldsymbol{P}) = \sum_{i=1}^{r}\sum_{j=r+1}^{m} \frac{\sigma_j^2}{\sigma_i^2 - \sigma_j^2}(\boldsymbol{v}_i\boldsymbol{v}_i^T) \otimes (\boldsymbol{u}_j\boldsymbol{u}_j^T)\,\mathrm{vec}(\boldsymbol{\Delta}) + \sum_{i=1}^{r}\sum_{j=r+1}^{m} \frac{\sigma_i\sigma_j}{\sigma_i^2 - \sigma_j^2}(\boldsymbol{v}_i\boldsymbol{u}_i^T) \otimes (\boldsymbol{u}_j\boldsymbol{v}_j^T)\,\mathrm{vec}(\boldsymbol{\Delta}^T) + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \quad (6.61)$$

Substituting (6.60) and (6.61) back into (6.54) yields

$$\mathrm{vec}\big(\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\big) = \sum_{i=1}^{r}\sum_{j=r+1}^{m} \Bigg(\frac{\sigma_j^2}{\sigma_i^2 - \sigma_j^2}\big((\boldsymbol{v}_j\boldsymbol{v}_j^T) \otimes (\boldsymbol{u}_i\boldsymbol{u}_i^T) + (\boldsymbol{v}_i\boldsymbol{v}_i^T) \otimes (\boldsymbol{u}_j\boldsymbol{u}_j^T)\big)\,\mathrm{vec}(\boldsymbol{\Delta}) + \frac{\sigma_i\sigma_j}{\sigma_i^2 - \sigma_j^2}\big((\boldsymbol{v}_j\boldsymbol{u}_j^T) \otimes (\boldsymbol{u}_i\boldsymbol{v}_i^T) + (\boldsymbol{v}_i\boldsymbol{u}_i^T) \otimes (\boldsymbol{u}_j\boldsymbol{v}_j^T)\big)\,\mathrm{vec}(\boldsymbol{\Delta}^T)\Bigg) + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2). \quad (6.62)$$

Truncating the inner summation, with $\sigma_j = 0$ for $j > n$, and applying Lemma 6.8-2

to the RHS of (6.62), we obtain

$$\boldsymbol{R_X}(\boldsymbol{\Delta}) = \sum_{i=1}^{r}\sum_{j=r+1}^{n}\Bigg(\frac{\sigma_j^2}{\sigma_i^2-\sigma_j^2}(\boldsymbol{u}_i\boldsymbol{u}_i^T\boldsymbol{\Delta}\boldsymbol{v}_j\boldsymbol{v}_j^T + \boldsymbol{u}_j\boldsymbol{u}_j^T\boldsymbol{\Delta}\boldsymbol{v}_i\boldsymbol{v}_i^T) + \frac{\sigma_i\sigma_j}{\sigma_i^2-\sigma_j^2}(\boldsymbol{u}_i\boldsymbol{v}_i^T\boldsymbol{\Delta}^T\boldsymbol{u}_j\boldsymbol{v}_j^T + \boldsymbol{u}_j\boldsymbol{v}_j^T\boldsymbol{\Delta}^T\boldsymbol{u}_i\boldsymbol{v}_i^T)\Bigg) + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2).$$

Our theorem now follows on the definition of $\boldsymbol{R_X}(\boldsymbol{\Delta})$ in (6.46).

6.8.2.1 Proof of Lemma 6.9

Using the fact from Proposition 6.2 that $\boldsymbol{P}_{\tilde{U}_2} = \boldsymbol{P}_{\hat{U}_2}$, we can re-express the subspace difference as

$$\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}} = \boldsymbol{P}_{\tilde{U}_2} - \boldsymbol{P}_{\boldsymbol{U}_2} = \boldsymbol{P}_{\hat{U}_2} - \boldsymbol{P}_{\boldsymbol{U}_2} = \hat{\boldsymbol{U}}_2\hat{\boldsymbol{U}}_2^T - \boldsymbol{U}_2\boldsymbol{U}_2^T. \tag{6.63}$$

Substituting (6.7b) into (6.63) yields

$$\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}} = (\boldsymbol{U}_2 + \boldsymbol{U}_1\boldsymbol{Q}^T)(\boldsymbol{I}_{m-r} + \boldsymbol{Q}\boldsymbol{Q}^T)^{-1}(\boldsymbol{U}_2^T + \boldsymbol{Q}\boldsymbol{U}_1^T) - \boldsymbol{U}_2\boldsymbol{U}_2^T. \tag{6.64}$$

Since $\boldsymbol{Q} = \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F)$ and $(\boldsymbol{I}_{m-r} + \boldsymbol{Q}\boldsymbol{Q}^T)^{-1} = \boldsymbol{I}_{m-r} - \boldsymbol{Q}\boldsymbol{Q}^T(\boldsymbol{I}_{m-r} + \boldsymbol{Q}\boldsymbol{Q}^T)^{-1} = \boldsymbol{I}_{m-r} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2)$, (6.64) can be simplified by absorbing second-order terms:

$$\begin{aligned}\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}} &= (\boldsymbol{U}_2 + \boldsymbol{U}_1\boldsymbol{Q}^T)(\boldsymbol{U}_2^T + \boldsymbol{Q}\boldsymbol{U}_1^T) - \boldsymbol{U}_2\boldsymbol{U}_2^T + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2) \\ &= \boldsymbol{U}_1\boldsymbol{Q}^T\boldsymbol{U}_2^T + \boldsymbol{U}_2\boldsymbol{Q}\boldsymbol{U}_1^T + \boldsymbol{U}_1\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{U}_1^T + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2) \\ &= \boldsymbol{U}_1\boldsymbol{Q}^T\boldsymbol{U}_2^T + \boldsymbol{U}_2\boldsymbol{Q}\boldsymbol{U}_1^T + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2).\end{aligned}$$

The equation $\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}} = -\boldsymbol{V}_1\boldsymbol{P}^T\boldsymbol{V}_2^T - \boldsymbol{V}_2\boldsymbol{P}\boldsymbol{V}_1^T + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^2)$ can be proved by a similar derivation. Since $\boldsymbol{Q}$ and $\boldsymbol{P}$ are in the order of $\|\boldsymbol{\Delta}\|_F$, so do $\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}$ and $\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}}$.

6.8.2.2 Proof of Lemma 6.10

Recall that

$$\boldsymbol{\Sigma}_1^2 = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ & \ddots & \\ 0 & \dots & \sigma_r^2 \end{bmatrix} \in \mathbb{R}^{r\times r} \qquad \text{and} \qquad \boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^T = \begin{bmatrix} \sigma_{r+1}^2 & \dots & 0 \\ & \ddots & \\ 0 & \dots & \sigma_m^2 \end{bmatrix} \in \mathbb{R}^{(m-r)\times(m-r)}.$$

By the definition of the Kronecker product, we have

$$\boldsymbol{I}_{m-r} \otimes \boldsymbol{\Sigma}_1^2 - \boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_2^T \otimes \boldsymbol{I}_r = \begin{bmatrix} \boldsymbol{\Sigma}_1^2 - \sigma_{r+1}^2\boldsymbol{I}_r & \dots & \boldsymbol{0}_r \\ & \ddots & \\ \boldsymbol{0}_r & \dots & \boldsymbol{\Sigma}_1^2 - \sigma_m^2\boldsymbol{I}_r \end{bmatrix} \in \mathbb{R}^{(m-r)r\times(m-r)r}.$$

Therefore, we can invert this diagonal matrix by considering each of the $r \times r$ blocks:

$$\begin{aligned}\boldsymbol{D} &= (\boldsymbol{I}_{m-r} \otimes \boldsymbol{\Sigma}_1^2 - \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_2^T \otimes \boldsymbol{I}_r)^{-1} \\ &= \begin{bmatrix} (\boldsymbol{\Sigma}_1^2 - \sigma_{r+1}^2 \boldsymbol{I}_r)^{-1} & \dots & \boldsymbol{0}_r \\ & \ddots & \\ \boldsymbol{0}_r & \dots & (\boldsymbol{\Sigma}_1^2 - \sigma_m^2 \boldsymbol{I}_r)^{-1} \end{bmatrix}.\end{aligned}$$

Now it is easy to verify that, for $i = 1, \dots, r$ and $k = 1, \dots, m-r$, the i-th diagonal entry of the k-th diagonal block, is $d_{ik} = 1/(\sigma_i^2 - \sigma_{r+k}^2)$. Furthermore, since $\big(\boldsymbol{e}_k^{m-r}(\boldsymbol{e}_k^{m-r})^T\big) \otimes \big(\boldsymbol{e}_i^r(\boldsymbol{e}_i^r)^T\big)$ is a $(m-r)r \times (m-r)r$ matrix of all zeros but the i-th diagonal entry of the k-th diagonal block is 1, we represent $\boldsymbol{D}$ as the sum of $(m-r)r$ rank-1 matrices:

$$\boldsymbol{D} = \sum_{i=1}^{r} \sum_{k=1}^{m-r} d_{ik} \big(\boldsymbol{e}_k^{m-r}(\boldsymbol{e}_k^{m-r})^T\big) \otimes \big(\boldsymbol{e}_i^r(\boldsymbol{e}_i^r)^T\big).$$

6.8.3 Proof of Theorem 6.2

By the definition of the r-TSVD in (6.2), we have

$$\mathcal{P}_r(\tilde{\boldsymbol{X}}) = \boldsymbol{P}_{\tilde{\boldsymbol{U}}_1} \tilde{\boldsymbol{X}} \boldsymbol{P}_{\tilde{\boldsymbol{V}}_1}. \tag{6.65}$$

Since we assume $\boldsymbol{X}$ has exact rank r, the perturbed matrix can be represented as $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T + \boldsymbol{\Delta}$. Substituting this back into (6.65) yields

$$\mathcal{P}_r(\boldsymbol{X} + \boldsymbol{\Delta}) = \boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}(\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T + \boldsymbol{\Delta})\boldsymbol{P}_{\tilde{\boldsymbol{V}}_1}. \tag{6.66}$$

Similar to the derivation of (6.64), we obtain $\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1} = (\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{I}_r + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}(\boldsymbol{U}_1^T - \boldsymbol{Q}^T\boldsymbol{U}_2^T)$ and $\boldsymbol{P}_{\tilde{\boldsymbol{V}}_1} = (\boldsymbol{V}_1 + \boldsymbol{V}_2\boldsymbol{P})(\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1}(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T)$. Substituting the expressions of $\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}$ and $\boldsymbol{P}_{\tilde{\boldsymbol{V}}_1}$ back into (6.66), we obtain

$$\begin{aligned}\mathcal{P}_r(\boldsymbol{X} + \boldsymbol{\Delta}) = (\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{I}_r + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}(\boldsymbol{U}_1^T - \boldsymbol{Q}^T\boldsymbol{U}_2^T)(\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T + \boldsymbol{\Delta}) \\ \cdot (\boldsymbol{V}_1 + \boldsymbol{V}_2\boldsymbol{P})(\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1}(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T).\end{aligned} \tag{6.67}$$

By orthogonality, the product of three terms in the middle of the RHS of (6.67) can be expanded and simplified as

$$\begin{aligned}&(\boldsymbol{U}_1^T - \boldsymbol{Q}^T\boldsymbol{U}_2^T)(\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T + \boldsymbol{\Delta})(\boldsymbol{V}_1 + \boldsymbol{V}_2\boldsymbol{P}) \\ &\qquad = (\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11}) + (\boldsymbol{E}_{12}\boldsymbol{P} - \boldsymbol{Q}^T\boldsymbol{E}_{21} - \boldsymbol{Q}^T\boldsymbol{E}_{22}\boldsymbol{P}).\end{aligned}$$

Therefore, (6.67) is equivalent to

$$\begin{aligned}\mathcal{P}_r(\boldsymbol{X} + \boldsymbol{\Delta}) = (\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{I}_r + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})(\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1}(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T) \\ + (\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{I}_r + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}(\boldsymbol{E}_{12}\boldsymbol{P} - \boldsymbol{Q}^T\boldsymbol{E}_{21} - \boldsymbol{Q}^T\boldsymbol{E}_{22}\boldsymbol{P}) \\ \cdot (\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1}(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T).\end{aligned} \tag{6.68}$$

Let us first focus on the first term on the RHS of (6.68). Similar to the result after (6.64), we have $(\boldsymbol{I}_r + \boldsymbol{Q}^T\boldsymbol{Q})^{-1} = \boldsymbol{I}_r - (\boldsymbol{I}_r + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{Q}$ and $(\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1} = \boldsymbol{I}_r - \boldsymbol{P}^T\boldsymbol{P}(\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1}$, and hence

$$
\begin{aligned}
&(\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{I}_r + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})(\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1} \\
&\quad = (\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})\Big(\boldsymbol{I}_r - (\boldsymbol{I}_r + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{Q}\Big)(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11}) \\
&\qquad \Big(\boldsymbol{I}_r - \boldsymbol{P}^T\boldsymbol{P}(\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1}\Big)(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T) \\
&\quad = (\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T) \\
&\qquad - (\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})\boldsymbol{P}^T\boldsymbol{P}(\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1}(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T) \\
&\quad - (\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{I}_r + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{Q}(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})(\boldsymbol{I}_r + \boldsymbol{P}^T\boldsymbol{P})^{-1}(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T).
\end{aligned}
\tag{6.69}
$$

Recall that $\boldsymbol{X} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T$ and $\boldsymbol{E}_{11} = \boldsymbol{U}_1^T\boldsymbol{\Delta}\boldsymbol{V}_1$. The product $(\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T)$ can be expanded as

$$
\begin{aligned}
(\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T) &= \boldsymbol{X} + \boldsymbol{U}_1\boldsymbol{E}_{11}\boldsymbol{V}_1^T \\
&+ \boldsymbol{U}_1(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})\boldsymbol{P}^T\boldsymbol{V}_2^T - \boldsymbol{U}_2\boldsymbol{Q}(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})\boldsymbol{V}_1^T - \boldsymbol{U}_2\boldsymbol{Q}(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})\boldsymbol{P}^T\boldsymbol{V}_2^T.
\end{aligned}
\tag{6.70}
$$

In order to make up the first-order terms that involve $\boldsymbol{\Delta}$, we need to decompose the perturbation into 4 components corresponding to different subspaces as follows.

Since $\boldsymbol{P}_{\boldsymbol{U}_1} + \boldsymbol{P}_{\boldsymbol{U}_2} = \boldsymbol{I}_m$ and $\boldsymbol{P}_{\boldsymbol{V}_1} + \boldsymbol{P}_{\boldsymbol{V}_2} = \boldsymbol{I}_n$, we have

$$\boldsymbol{\Delta} = \boldsymbol{P}_{\boldsymbol{U}_1}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_1} + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_1} + \boldsymbol{P}_{\boldsymbol{U}_1}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}. \tag{6.71}$$

Reorganizing terms in (6.71) as

$$\boldsymbol{P}_{\boldsymbol{U}_1}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_1} = \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} - \boldsymbol{P}_{\boldsymbol{U}_1}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_1},$$

and using the definition of $\boldsymbol{E}$ in (6.4), we further have

$$\boldsymbol{U}_1\boldsymbol{E}_{11}\boldsymbol{V}_1^T = \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} - \boldsymbol{U}_1\boldsymbol{E}_{12}\boldsymbol{V}_2^T - \boldsymbol{U}_2\boldsymbol{E}_{21}\boldsymbol{V}_1^T. \tag{6.72}$$

Thus, substituting (6.72) back into (6.70) and rearranging terms yield

$$\begin{aligned}(\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})(\boldsymbol{V}_1^T + \boldsymbol{P}^T\boldsymbol{V}_2^T)& \\ = \boldsymbol{X} + \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{U}_1\big((\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})\boldsymbol{P}^T - \boldsymbol{E}_{12}\big)\boldsymbol{V}_2^T& \\ - \boldsymbol{U}_2\big(\boldsymbol{Q}(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11}) + \boldsymbol{E}_{21}\big)\boldsymbol{V}_1^T - \boldsymbol{U}_2\boldsymbol{Q}(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})\boldsymbol{P}^T\boldsymbol{V}_2^T.&\end{aligned} \tag{6.73}$$

Substituting (6.69) and (6.73) back into (6.68), we obtain

$$\begin{aligned}\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) &= \boldsymbol{X}+\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}+\boldsymbol{U}_1\big((\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\boldsymbol{P}^T-\boldsymbol{E}_{12}\big)\boldsymbol{V}_2^T\\ &-\boldsymbol{U}_2\big(\boldsymbol{Q}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})+\boldsymbol{E}_{21}\big)\boldsymbol{V}_1^T-\boldsymbol{U}_2\boldsymbol{Q}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\boldsymbol{P}^T\boldsymbol{V}_2^T\\ &+(\boldsymbol{U}_1-\boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{I}_r+\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\Big(-(\boldsymbol{I}_r+\boldsymbol{Q}^T\boldsymbol{Q})(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\boldsymbol{P}^T\boldsymbol{P}+\boldsymbol{Q}^T\boldsymbol{Q}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\\ &\qquad+(\boldsymbol{E}_{12}\boldsymbol{P}-\boldsymbol{Q}^T\boldsymbol{E}_{21}-\boldsymbol{Q}^T\boldsymbol{E}_{22}\boldsymbol{P})\Big)(\boldsymbol{I}_r+\boldsymbol{P}^T\boldsymbol{P})^{-1}(\boldsymbol{V}_1^T+\boldsymbol{P}^T\boldsymbol{V}_2^T). \qquad (6.74)\end{aligned}$$

Applying (6.6), we have

$$\begin{aligned}&\boldsymbol{U}_1\big((\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\boldsymbol{P}^T-\boldsymbol{E}_{12}\big)\boldsymbol{V}_2^T-\boldsymbol{U}_2\big(\boldsymbol{Q}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})+\boldsymbol{E}_{21}\big)\boldsymbol{V}_1^T\\ &\qquad-\boldsymbol{U}_2\boldsymbol{Q}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\boldsymbol{P}^T\boldsymbol{V}_2^T\\ &=\boldsymbol{U}_1\boldsymbol{Q}^T(\boldsymbol{E}_{21}\boldsymbol{P}^T-\boldsymbol{E}_{22})\boldsymbol{V}_2^T+\boldsymbol{U}_2(\boldsymbol{E}_{22}-\boldsymbol{Q}\boldsymbol{E}_{12})\boldsymbol{P}\boldsymbol{V}_1^T\\ &\qquad+\boldsymbol{U}_2(\boldsymbol{E}_{21}+\boldsymbol{E}_{22}\boldsymbol{P}+\boldsymbol{Q}\boldsymbol{E}_{12}\boldsymbol{P})\boldsymbol{P}^T\boldsymbol{V}_2^T, \qquad (6.75)\end{aligned}$$

and

$$\begin{aligned}&-(\boldsymbol{I}_r+\boldsymbol{Q}^T\boldsymbol{Q})(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\boldsymbol{P}^T\boldsymbol{P}+\boldsymbol{Q}^T\boldsymbol{Q}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})+(\boldsymbol{E}_{12}\boldsymbol{P}-\boldsymbol{Q}^T\boldsymbol{E}_{21}-\boldsymbol{Q}^T\boldsymbol{E}_{22}\boldsymbol{P})\\ &=\big(\boldsymbol{E}_{12}\boldsymbol{P}-(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\boldsymbol{P}^T\boldsymbol{P}\big)-\big(\boldsymbol{Q}^T\boldsymbol{E}_{21}+\boldsymbol{Q}^T\boldsymbol{Q}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\big)\\ &\qquad+\boldsymbol{Q}^T\big(\boldsymbol{E}_{22}+\boldsymbol{Q}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\boldsymbol{P}^T\big)\boldsymbol{P}\\ &=(\boldsymbol{Q}^T\boldsymbol{E}_{22}-\boldsymbol{Q}^T\boldsymbol{E}_{21}\boldsymbol{P}^T)\boldsymbol{P}-\boldsymbol{Q}^T(\boldsymbol{E}_{22}\boldsymbol{P}+\boldsymbol{Q}\boldsymbol{E}_{12}\boldsymbol{P})+\boldsymbol{Q}^T\big(\boldsymbol{E}_{22}\\ &\qquad+\boldsymbol{Q}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\boldsymbol{P}^T\big)\boldsymbol{P}. \qquad (6.76)\end{aligned}$$

Substituting (6.75) and (6.76) back into (6.74), we obtain

$$\begin{aligned}\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) &= \boldsymbol{X}+\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}+\boldsymbol{U}_1\boldsymbol{Q}^T(\boldsymbol{E}_{21}\boldsymbol{P}^T-\boldsymbol{E}_{22})\boldsymbol{V}_2^T \\ &+\boldsymbol{U}_2(\boldsymbol{E}_{22}-\boldsymbol{Q}\boldsymbol{E}_{12})\boldsymbol{P}\boldsymbol{V}_1^T+\boldsymbol{U}_2(\boldsymbol{E}_{21}+\boldsymbol{E}_{22}\boldsymbol{P}+\boldsymbol{Q}\boldsymbol{E}_{12}\boldsymbol{P})\boldsymbol{P}^T\boldsymbol{V}_2^T \\ &+(\boldsymbol{U}_1-\boldsymbol{U}_2\boldsymbol{Q})(\boldsymbol{I}_r+\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\cdot\Big((\boldsymbol{Q}^T\boldsymbol{E}_{22}-\boldsymbol{Q}^T\boldsymbol{E}_{21}\boldsymbol{P}^T)\boldsymbol{P}-\boldsymbol{Q}^T(\boldsymbol{E}_{22}\boldsymbol{P}+\boldsymbol{Q}\boldsymbol{E}_{12}\boldsymbol{P}) \\ &\qquad+\boldsymbol{Q}^T\big(\boldsymbol{E}_{22}+\boldsymbol{Q}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})\boldsymbol{P}^T\big)\boldsymbol{P}\Big)(\boldsymbol{I}_r+\boldsymbol{P}^T\boldsymbol{P})^{-1}(\boldsymbol{V}_1^T+\boldsymbol{P}^T\boldsymbol{V}_2^T).\end{aligned} \tag{6.77}$$

Since $\boldsymbol{Q}, \boldsymbol{P}, \boldsymbol{E}_{11}, \boldsymbol{E}_{12}, \boldsymbol{E}_{21}$, and $\boldsymbol{E}_{22}$ are first-order, and $(\boldsymbol{I}_r+\boldsymbol{Q}^T\boldsymbol{Q})^{-1}$, $(\boldsymbol{I}_r+\boldsymbol{P}^T\boldsymbol{P})^{-1}$ are zero-order in terms of $\|\boldsymbol{\Delta}\|_F$, we can collect all the third-order terms on the RHS of (6.77) and obtain

$$\begin{aligned}\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) &= \boldsymbol{X}+\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}-\boldsymbol{U}_1\boldsymbol{Q}^T\boldsymbol{E}_{22}\boldsymbol{V}_2^T+\boldsymbol{U}_2\boldsymbol{E}_{22}\boldsymbol{P}\boldsymbol{V}_1^T \\ &\quad+\boldsymbol{U}_2\boldsymbol{E}_{21}\boldsymbol{P}^T\boldsymbol{V}_2^T+\boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^3).\end{aligned} \tag{6.78}$$

Finally, the matrices $\boldsymbol{Q}$ and $\boldsymbol{P}$ in the second-order terms is eliminated by the following variant of (6.6):

$$\begin{aligned}\boldsymbol{Q} &= -\big(\boldsymbol{E}_{21}+\boldsymbol{Q}\boldsymbol{E}_{21}\boldsymbol{P}-\boldsymbol{E}_{22}\boldsymbol{P}-\boldsymbol{Q}\boldsymbol{E}_{11}\big)\boldsymbol{\Sigma}_1^{-1}, \\ \boldsymbol{P}^T &= \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{E}_{12}+\boldsymbol{Q}^T\boldsymbol{E}_{21}\boldsymbol{P}^T-\boldsymbol{Q}^T\boldsymbol{E}_{22}-\boldsymbol{E}_{11}\boldsymbol{P}^T).\end{aligned}$$

The substitution and collection of third-order terms on the RHS of (6.78) yield

$$\begin{aligned}
\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) &= \boldsymbol{X}+\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}+\boldsymbol{U}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{E}_{21}^T\boldsymbol{E}_{22}\boldsymbol{V}_2^T+\boldsymbol{U}_2\boldsymbol{E}_{22}\boldsymbol{E}_{12}^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{V}_1^T \\
&\quad +\boldsymbol{U}_2\boldsymbol{E}_{21}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{E}_{12}\boldsymbol{V}_2^T+\boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^3) \\
&= \boldsymbol{X}+\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}+\boldsymbol{U}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{V}_1^T\boldsymbol{\Delta}^T\boldsymbol{U}_2\boldsymbol{U}_2^T\boldsymbol{\Delta}\boldsymbol{V}_2\boldsymbol{V}_2^T \\
&\quad +\boldsymbol{U}_2\boldsymbol{U}_2^T\boldsymbol{\Delta}\boldsymbol{V}_2\boldsymbol{V}_2^T\boldsymbol{\Delta}^T\boldsymbol{U}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{V}_1^T+\boldsymbol{U}_2\boldsymbol{U}_2^T\boldsymbol{\Delta}\boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^T\boldsymbol{\Delta}\boldsymbol{V}_2\boldsymbol{V}_2^T+\boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^3) \\
&= \boldsymbol{X}+\boldsymbol{\Delta}-\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}+\boldsymbol{X}^\dagger\boldsymbol{\Delta}^T P_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}+\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^T\boldsymbol{X}^\dagger \\
&\quad +\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^\dagger)^T\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}+\boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^3).
\end{aligned}$$

This completes our proof of the theorem.

6.8.4 Proof of Lemma 6.1

By the triangle inequality, we have

$$\|\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta})-(\boldsymbol{X}+\boldsymbol{\Delta})+\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\|_F \le \|\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta})-(\boldsymbol{X}+\boldsymbol{\Delta})\|_F+\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\|_F\,. \tag{6.79}$$

The first term on the RHS of (6.79) can be bounded as follows. Since $\tilde{\boldsymbol{X}}=\boldsymbol{X}+\boldsymbol{\Delta}$, applying the norm absolute homogeneity property yields

$$\|\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta})-(\boldsymbol{X}+\boldsymbol{\Delta})\|_F=\left\|\mathcal{P}_r(\tilde{\boldsymbol{X}})-\tilde{\boldsymbol{X}}\right\|_F=\left\|\tilde{\boldsymbol{X}}-\mathcal{P}_r(\tilde{\boldsymbol{X}})\right\|_F. \tag{6.80}$$

From Lemmas 6.4 and 6.7, we obtain

$$\left\|\tilde{\boldsymbol{X}} - \mathcal{P}_r(\tilde{\boldsymbol{X}})\right\|_F = \left\|\tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T\right\|_F = \left\|\tilde{\boldsymbol{\Sigma}}_2\right\|_F. \tag{6.81}$$

Since $\tilde{\boldsymbol{\Sigma}}_2$ is a submatrix of $\tilde{\boldsymbol{\Sigma}}$ containing $n-r$ small singular values of $\tilde{\boldsymbol{X}}$ in the diagonal, it holds that

$$\left\|\tilde{\boldsymbol{\Sigma}}_2\right\|_F \leq \left\|\tilde{\boldsymbol{\Sigma}}\right\|_F = \left\|\tilde{\boldsymbol{X}}\right\|_F. \tag{6.82}$$

Additionally, using the triangle inequality we can bound $\left\|\tilde{\boldsymbol{X}}\right\|_F$ by

$$\left\|\tilde{\boldsymbol{X}}\right\|_F = \|\boldsymbol{X}+\boldsymbol{\Delta}\|_F \leq \|\boldsymbol{X}\|_F + \|\boldsymbol{\Delta}\|_F. \tag{6.83}$$

From (6.80), (6.81), (6.82), and (6.83), we have

$$\|\mathcal{P}_r(\boldsymbol{X}+\boldsymbol{\Delta}) - (\boldsymbol{X}+\boldsymbol{\Delta})\|_F \leq \|\boldsymbol{X}\|_F + \|\boldsymbol{\Delta}\|_F. \tag{6.84}$$

On the other hand, it follows from Lemma 6.7 that the second term on the RHS of (6.79) satisfies

$$\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\|_F \leq \|\boldsymbol{\Delta}\|_F. \tag{6.85}$$

Substituting inequalities (6.84) and (6.85) into (6.79) completes the proof of the lemma.

6.8.5 Proof of Theorem 6.3

The following proof is developed for the case of a rank-r matrix $\boldsymbol{X}$. We first derive the proof of (6.28) and then use this result to prove (6.29).

6.8.5.1 Proof of the bound in (6.28)

Our goal is to prove that the residual in (6.27) is always bounded by

$$\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F \leq \frac{c}{\sigma_r}\|\boldsymbol{\Delta}\|_F^2, \qquad \text{for some } 1+1/\sqrt{2} \leq c \leq 4(1+\sqrt{2}).$$

Let us begin with the upper bound on c by showing that

$$\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F \leq \frac{4(1+\sqrt{2})}{\sigma_r}\|\boldsymbol{\Delta}\|_F^2. \tag{6.86}$$

Rearranging terms in (6.27) and replacing $\boldsymbol{X}+\boldsymbol{\Delta}$ by $\tilde{\boldsymbol{X}}$, we have

$$\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) = \mathcal{P}_r(\tilde{\boldsymbol{X}}) - \tilde{\boldsymbol{X}} + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}. \tag{6.87}$$

Using the singular subspace decomposition in Definition 6.2 with descending order of singular values $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \ldots \geq \tilde{\sigma}_n$, let us decompose $\tilde{\boldsymbol{X}}$ as follows

$$\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{\Sigma}}_1\tilde{\boldsymbol{V}}_1^T + \tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T. \tag{6.88}$$

Since in this theorem we consider perturbations of any magnitude, $\tilde{\boldsymbol{X}}$ can take any value including the case in which $\tilde{\sigma}_r = \tilde{\sigma}_{r+1}$ and the decomposition (6.88) may not be unique. Nevertheless, the proof holds for any valid choice of singular subspace decomposition. From such a choice in (6.88), $\mathcal{P}_r(\tilde{\boldsymbol{X}})$ is well-defined as: $\mathcal{P}_r(\tilde{\boldsymbol{X}}) = \tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{\Sigma}}_1\tilde{\boldsymbol{V}}_1^T$. Substituting $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta}$ into (6.48) and using the fact that $\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{X} = \boldsymbol{0}$ and $\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{V}_2} = \boldsymbol{0}$, we obtain

$$\begin{aligned}\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) &= -\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} - \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} - \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} \\ &= -\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} - \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2}. \end{aligned} \tag{6.89}$$

Here, from Lemma 6.3, we can replace $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^\dagger)^T\boldsymbol{X}$ in the first term on the RHS of (6.89) and obtain

$$\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) = -(\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{X})(\boldsymbol{X}^\dagger)^T(\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}) - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} - \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2}. \tag{6.90}$$

Taking the Frobenius norm and using its absolute homogeneity property, (6.90) becomes

$$\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F = \left\|(\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{X})(\boldsymbol{X}^\dagger)^T(\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}) + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} + \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2}\right\|_F.$$

By the triangle inequality, the norm of $\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})$ is then bounded by

$$\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F \leq \left\|(\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{X})(\boldsymbol{X}^\dagger)^T(\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}})\right\|_F + \left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\right\|_F + \left\|\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2}\right\|_F. \tag{6.91}$$

Let us proceed to upper-bound $\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F$ by finding the upper bounds for each of the three terms on the RHS of (6.91) with respect to $\|\boldsymbol{\Delta}\|_F^2$. Our proof technique utilizes the following lemmas.

Lemma 6.11. $\max\{\|\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}\boldsymbol{X}\|_F, \|\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}}\|_F\} \leq 2\|\boldsymbol{\Delta}\|_F$.

Lemma 6.12. $\max\{\|\boldsymbol{P}_{U_2}\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}\boldsymbol{\Delta}\|_F, \|\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}}\boldsymbol{P}_{V_2}\|_F\} \leq \frac{2}{\sigma_r}\|\boldsymbol{\Delta}\|_F^2$.

The proofs of Lemmas 6.11 and 6.12 are given at the end of this subsection. Let us proceed with the task of bounding the first term in (6.91). Applying Lemma 6.5 twice and using the fact that $\|\boldsymbol{X}^\dagger\|_2 = 1/\sigma_r$, we have

$$\|(\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}\boldsymbol{X})(\boldsymbol{X}^\dagger)^T(\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}})\|_F \leq \frac{1}{\sigma_r}\|\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}\boldsymbol{X}\|_F\|\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}}\|_F. \tag{6.92}$$

By Lemma 6.11, the terms $\|\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}\boldsymbol{X}\|_F$ and $\|\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}}\|_F$ can each be bounded by $2\|\boldsymbol{\Delta}\|_F$. Applying the upper bounds on the RHS of (6.92), we obtain the following bound on the first term in (6.91):

$$\|(\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}\boldsymbol{X})(\boldsymbol{X}^\dagger)^T(\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}})\|_F \leq \frac{4}{\sigma_r}\|\boldsymbol{\Delta}\|_F^2. \tag{6.93}$$

Next, we shall bound the second term in (6.91), i.e., $\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}}\|_F$. From Lemma 6.7, we have

$$\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}}\|_F \leq \|\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{V_2}}\|_F. \tag{6.94}$$

Since $\boldsymbol{P}_{\boldsymbol{V}_1} + \boldsymbol{P}_{\boldsymbol{V}_2} = \boldsymbol{I}_n$, the matrix on the RHS of (6.94) can be expanded as the

sum of two orthogonal terms:

$$\boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} = \boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}(\boldsymbol{P}_{\boldsymbol{V}_1} + \boldsymbol{P}_{\boldsymbol{V}_2}) = \boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_1} + \boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_2}.$$

Notice that $\boldsymbol{P}_{\boldsymbol{V}_1}$ and $\boldsymbol{P}_{\boldsymbol{V}_2}$ are orthogonal. By Lemma 6.6, we have

$$\begin{aligned}
\left\|\boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\right\|_F^2 &= \left\|\boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_1}\right\|_F^2 + \left\|\boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_2}\right\|_F^2 \\
&= \left\|\boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{X}^T\boldsymbol{X}^\dagger\right\|_F^2 + \left\|\boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_2}\right\|_F^2 \qquad (\text{since } \boldsymbol{P}_{\boldsymbol{V}_1} = \boldsymbol{X}^T\boldsymbol{X}^\dagger) \\
&= \left\|\boldsymbol{\Delta}(\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}})^T\boldsymbol{X}^\dagger\right\|_F^2 + \left\|\boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_2}\right\|_F^2. \qquad (6.95)
\end{aligned}$$

Each term on the RHS of (6.95) can be bounded as follows. Applying Lemma 6.5 twice, we initially bound the first term on the RHS of (6.95) as follows:

$$\left\|\boldsymbol{\Delta}(\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}})^T\boldsymbol{X}^\dagger\right\|_F \leq \frac{1}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F\left\|\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\right\|_F.$$

By Lemma 6.11, we upper-bound $\left\|\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\right\|_F$ by $2\left\|\boldsymbol{\Delta}\right\|_F$ and obtain the bound on the first term on the RHS of (6.95):

$$\left\|\boldsymbol{\Delta}(\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}})^T\boldsymbol{X}^\dagger\right\|_F \leq \frac{2}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F^2. \qquad (6.96)$$

To bound the second term on the RHS of (6.95), we apply Lemma 6.12 and obtain

$$\left\|\boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_2}\right\|_F^2 \leq \frac{4}{\sigma_r^2}\left\|\boldsymbol{\Delta}\right\|_F^4. \qquad (6.97)$$

Substituting the bounds from (6.96) and (6.97) back into the RHS of (6.95), we have

$$\left\|\boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\right\|_F^2 \leq \left(\frac{2}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F^2\right)^2 + \frac{4}{\sigma_r^2}\left\|\boldsymbol{\Delta}\right\|_F^4 = \frac{8}{\sigma_r^2}\left\|\boldsymbol{\Delta}\right\|_F^4.$$

Taking the square root of the last result and substituting it back to (6.94) yields

$$\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\right\|_F \leq \frac{2\sqrt{2}}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F^2. \tag{6.98}$$

This offers a bound on the second term on the RHS of (6.91). Similarly, we bound the third term on the RHS of (6.91) by

$$\left\|\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta P}_{\check{\boldsymbol{V}}_2}\right\|_F \leq \frac{2\sqrt{2}}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F^2. \tag{6.99}$$

Finally, summing up (6.93), (6.98), and (6.99), and substituting back into (6.91), we obtain (6.86) and thereby completes the first part of the proof.

For the second part of the proof, we show that $c \geq 1 + 1/\sqrt{2}$ by constructing a perturbation $\boldsymbol{\Delta}$ such that the ratio $\left\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\right\|_F / \left\|\boldsymbol{\Delta}\right\|_F^2$ approaches $(1+1/\sqrt{2})/\sigma_r$. Consider perturbations of form

$$\boldsymbol{\Delta} = (\sigma - \sigma_r - \epsilon)\boldsymbol{u}_r\boldsymbol{v}_r^T + \sigma\boldsymbol{u}_{r+1}\boldsymbol{v}_{r+1}^T, \qquad \text{for } 0 < \epsilon < \sigma < \sigma_r. \tag{6.100}$$

Since $\boldsymbol{u}_r\boldsymbol{v}_r^T$ and $\boldsymbol{u}_{r+1}\boldsymbol{v}_{r+1}^T$ are orthogonal, we can compute the norm of $\boldsymbol{\Delta}$ using

Lemma 6.6:

$$\begin{aligned}\|\boldsymbol{\Delta}\|_F^2 &= (\sigma - \sigma_r - \epsilon)^2 \left\|\boldsymbol{u}_r\boldsymbol{v}_r^T\right\|_F^2 + \sigma^2 \left\|\boldsymbol{u}_{r+1}\boldsymbol{v}_{r+1}^T\right\|_F^2 \\ &= (\sigma - \sigma_r - \epsilon)^2 + \sigma^2, \end{aligned} \tag{6.101}$$

where the second equality uses $\boldsymbol{u}_r\boldsymbol{v}_r^T = \boldsymbol{u}_r \otimes \boldsymbol{v}_r^T$ and Lemma 6.8-3. Using the SVD of $\boldsymbol{X}$ and the definition of $\boldsymbol{\Delta}$ in (6.100), we have

$$\begin{aligned}\boldsymbol{X} + \boldsymbol{\Delta} &= \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T + (\sigma - \sigma_r - \epsilon)\boldsymbol{u}_r\boldsymbol{v}_r^T + \sigma\boldsymbol{u}_{r+1}\boldsymbol{v}_{r+1}^T \\ &= \sum_{i=1}^{r-1} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T + (\sigma - \epsilon)\boldsymbol{u}_r\boldsymbol{v}_r^T + \sigma\boldsymbol{u}_{r+1}\boldsymbol{v}_{r+1}^T. \end{aligned} \tag{6.102}$$

After perturbation, the r-th singular value of $\boldsymbol{X}$ is changed from σ_r to $\sigma - \epsilon$ and the $r+1$-th changes from 0 to σ, thereby making the singular value corresponding to $\boldsymbol{u}_{r+1}\boldsymbol{v}_{r+1}^T$ larger than the singular value associated with $\boldsymbol{u}_r\boldsymbol{v}_r^T$. Thus, the r-TSVD of $\boldsymbol{X} + \boldsymbol{\Delta}$ is given by

$$\mathcal{P}_r(\boldsymbol{X} + \boldsymbol{\Delta}) = \sum_{i=1}^{r-1} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T + \sigma\boldsymbol{u}_{r+1}\boldsymbol{v}_{r+1}^T. \tag{6.103}$$

On the other hand, since $\boldsymbol{P}_{\boldsymbol{U}_2} = \sum_{i=r+1}^{m} \boldsymbol{u}_i\boldsymbol{u}_i^T$ and $\boldsymbol{P}_{\boldsymbol{V}_2} = \sum_{i=r+1}^{n} \boldsymbol{v}_i\boldsymbol{v}_i^T$, we have

$$\begin{aligned}\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} &= \Big(\sum_{i=r+1}^{m} \boldsymbol{u}_i\boldsymbol{u}_i^T\Big)\Big((\sigma - \sigma_r - \epsilon)\boldsymbol{u}_r\boldsymbol{v}_r^T + \sigma\boldsymbol{u}_{r+1}\boldsymbol{v}_{r+1}^T\Big)\Big(\sum_{i=r+1}^{n} \boldsymbol{v}_i\boldsymbol{v}_i^T\Big) \\ &= \sigma\boldsymbol{u}_{r+1}\boldsymbol{v}_{r+1}^T, \end{aligned} \tag{6.104}$$

where the second equality stems from the fact that

$$\boldsymbol{u}_i^T \boldsymbol{u}_j = \boldsymbol{v}_i^T \boldsymbol{v}_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Substituting (6.102), (6.103), and (6.104) into (6.87), we obtain

$$\begin{aligned} \boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) &= \left(\sum_{i=1}^{r-1} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T + \sigma \boldsymbol{u}_{r+1} \boldsymbol{v}_{r+1}^T \right) \\ &\qquad - \left(\sum_{i=1}^{r-1} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T + (\sigma - \epsilon) \boldsymbol{u}_r \boldsymbol{v}_r^T + \sigma \boldsymbol{u}_{r+1} \boldsymbol{v}_{r+1}^T \right) + \sigma \boldsymbol{u}_{r+1} \boldsymbol{v}_{r+1}^T \\ &= -(\sigma - \epsilon) \boldsymbol{u}_r \boldsymbol{v}_r^T + \sigma \boldsymbol{u}_{r+1} \boldsymbol{v}_{r+1}^T. \end{aligned}$$

Similar to (6.101), one can compute the norm of the residual by

$$\left\| \boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) \right\|_F = \sqrt{(\sigma - \epsilon)^2 + \sigma^2}. \tag{6.105}$$

From (6.101) and (6.105), we have

$$\frac{\left\| \boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) \right\|_F}{\left\| \boldsymbol{\Delta} \right\|_F^2} = \frac{\sqrt{(\sigma - \epsilon)^2 + \sigma^2}}{(\sigma_r + \epsilon - \sigma)^2 + \sigma^2}.$$

Now maximizing over σ while taking ϵ to 0 gives us a lower bound on c:

$$\begin{aligned}\frac{c}{\sigma_r} &= \sup_{\boldsymbol{\Delta}\in\mathbb{R}^{m\times n}} \frac{\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2} \\ &\geq \max_{0<\sigma<\sigma_r} \lim_{\epsilon\to 0^+} \frac{\sqrt{(\sigma-\epsilon)^2+\sigma^2}}{(\sigma_r+\epsilon-\sigma)^2+\sigma^2} \\ &= \max_{0<\sigma<\sigma_r} \frac{\sigma\sqrt{2}}{(\sigma_r-\sigma)^2+\sigma^2}. \end{aligned} \tag{6.106}$$

The maximization can be obtained at $\sigma = \sigma_r/\sqrt{2}$. Therefore, substituting back into (6.106) yields $c \geq 1 + 1/\sqrt{2}$. This completes our proof of the first half of Theorem 6.3. We recall from Remark 6.3 that we conjecture the structure of $\boldsymbol{\Delta}$ given in (6.100) yields the maximizer of $\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F / \|\boldsymbol{\Delta}\|_F^2$.

Proof of Lemma 6.11 Let us rewrite $\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}\boldsymbol{X} = \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}\boldsymbol{X} - \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{X}$. Since $\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{X} = \boldsymbol{0}$, we obtain

$$\begin{aligned}\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}\boldsymbol{X} &= \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}\boldsymbol{X} \\ &= \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}(\tilde{\boldsymbol{X}} - \boldsymbol{\Delta}) \qquad\qquad (\text{since } \tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta}) \\ &= \tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{U}}_2^T\tilde{\boldsymbol{X}} - \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}\boldsymbol{\Delta}. \end{aligned} \tag{6.107}$$

Substituting $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{\Sigma}}_1\tilde{\boldsymbol{V}}_1^T + \tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T$ yields

$$\begin{aligned}\boldsymbol{\delta}_{\boldsymbol{P}_{U_2}}\boldsymbol{X} &= \tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{U}}_2^T\big(\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{\Sigma}}_1\tilde{\boldsymbol{V}}_1^T + \tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T\big) - \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}\boldsymbol{\Delta} \\ &= \tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T - \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}\boldsymbol{\Delta}, \end{aligned}$$

where in the last equality we use the fact that $\tilde{\boldsymbol{U}}_2^T\tilde{\boldsymbol{U}}_1 = \boldsymbol{0}$ and $\tilde{\boldsymbol{U}}_2^T\tilde{\boldsymbol{U}}_2 = \boldsymbol{I}_m$. Therefore,

$$\left\|\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{X}\right\|_F = \left\|\tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T - \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}\boldsymbol{\Delta}\right\|_F. \tag{6.108}$$

By the triangle inequality and the absolute homogeneity, (6.108) implies

$$\left\|\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{X}\right\|_F \leq \left\|\tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T\right\|_F + \left\|\boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}\boldsymbol{\Delta}\right\|_F. \tag{6.109}$$

We shall bound each term on the RHS of (6.109) as follows. First, using Lemma 6.7, we can remove the semi-orthogonal matrices from within the Frobenius norm without changing the value of the norm:

$$\left\|\tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T\right\|_F = \left\|\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T\right\|_F = \left\|\tilde{\boldsymbol{\Sigma}}_2\right\|_F.$$

Since $\boldsymbol{\Sigma}_2 = \boldsymbol{0}$, we further obtain

$$\left\|\tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T\right\|_F = \left\|\tilde{\boldsymbol{\Sigma}}_2 - \boldsymbol{\Sigma}_2\right\|_F. \tag{6.110}$$

In addition, recall that $\tilde{\boldsymbol{\Sigma}}_2$ and $\boldsymbol{\Sigma}_2$ are sub-matrices of $\tilde{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$, respectively. Thus,

$$\left\|\tilde{\boldsymbol{\Sigma}}_2 - \boldsymbol{\Sigma}_2\right\|_F \leq \left\|\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F. \tag{6.111}$$

Moreover, by Mirsky's inequality in Proposition 6.1, we have

$$\left\|\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F = \sqrt{\sum_{i=1}^{n}(\tilde{\sigma}_i - \sigma_i)^2} \leq \|\boldsymbol{\Delta}\|_F . \tag{6.112}$$

From (6.110), (6.111), and (6.112), it follows that

$$\left\|\tilde{\boldsymbol{U}}_2\tilde{\boldsymbol{\Sigma}}_2\tilde{\boldsymbol{V}}_2^T\right\|_F \leq \|\boldsymbol{\Delta}\|_F . \tag{6.113}$$

Next, the second term on the RHS of (6.109), by Lemma 6.7, is bounded by

$$\left\|\boldsymbol{P}_{\tilde{\boldsymbol{U}}_2}\boldsymbol{\Delta}\right\|_F \leq \|\boldsymbol{\Delta}\|_F . \tag{6.114}$$

Summing up (6.113) and (6.114), and combining the resulting inequality with (6.109), we conclude that

$$\left\|\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{X}\right\|_F \leq 2\|\boldsymbol{\Delta}\|_F .$$

The proof of $\left\|\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\right\|_F \leq 2\|\boldsymbol{\Delta}\|_F$ follows a similar derivation.

Proof of Lemma 6.12 In this subsection, we shall show that $\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\right\|_F \leq \frac{2}{\sigma_r}\|\boldsymbol{\Delta}\|_F^2$. The proof of $\left\|\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_2}\right\|_F \leq \frac{2}{\sigma_r}\|\boldsymbol{\Delta}\|_F^2$ can be derived similarly. Since

Definition 6.3 implies $\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}} = \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2} - \boldsymbol{P}_{\boldsymbol{U}_2} = \boldsymbol{P}_{\boldsymbol{U}_1} - \boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}$, we have

$$\begin{aligned} \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta} &= \boldsymbol{P}_{\boldsymbol{U}_2}(\boldsymbol{P}_{\boldsymbol{U}_1} - \boldsymbol{P}_{\tilde{\boldsymbol{U}}_1})\boldsymbol{\Delta} \\ &= -\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}\boldsymbol{\Delta}, \end{aligned} \tag{6.115}$$

where the second equality is due to $\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\boldsymbol{U}_1} = \boldsymbol{0}$ (see Lemma 6.3). It is now sufficient to bound the norm of $\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}\boldsymbol{\Delta}$ by $\frac{2}{\sigma_r}\|\boldsymbol{\Delta}\|_F^2$. Let us consider two cases:

- If $\|\boldsymbol{\Delta}\|_2 \geq \sigma_r/2$, then applying Lemma 6.7-2 twice yields

$$\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}\boldsymbol{\Delta}\right\|_F \leq \|\boldsymbol{\Delta}\|_F . \tag{6.116}$$

  Since $\|\boldsymbol{\Delta}\|_F \geq \|\boldsymbol{\Delta}\|_2 \geq \sigma_r/2$, multiplying both sides by $\frac{2}{\sigma_r}\|\boldsymbol{\Delta}\|_F$ yields

$$\|\boldsymbol{\Delta}\|_F \leq \frac{2}{\sigma_r}\|\boldsymbol{\Delta}\|_F^2 . \tag{6.117}$$

  From (6.116) and (6.117), we obtain $\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}\boldsymbol{\Delta}\right\|_F \leq \frac{2}{\sigma_r}\|\boldsymbol{\Delta}\|_F^2$.

- If $\|\boldsymbol{\Delta}\|_2 < \sigma_r/2$, we need to use a different approach as follows. First, from Lemma 6.5, we have

$$\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}\boldsymbol{\Delta}\right\|_F \leq \left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}\right\|_2 \|\boldsymbol{\Delta}\|_F . \tag{6.118}$$

  Let us examine the product $\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}$. Let $\tilde{\boldsymbol{X}}_1 = \tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{\Sigma}}_1\tilde{\boldsymbol{V}}_1^T$ and $\tilde{\boldsymbol{X}}_2 = \tilde{\boldsymbol{X}} - \tilde{\boldsymbol{X}}_1$.

From Weyl's inequality [226], we have

$$|\tilde{\sigma}_i - \sigma_i| \leq \|\boldsymbol{\Delta}\|_2 < \frac{\sigma_r}{2} \qquad \text{for } i = 1, \ldots, n.$$

Thus, for any $1 \leq i \leq r$, it holds that

$$\tilde{\sigma}_i > \sigma_i - \frac{\sigma_r}{2} \geq \sigma_r - \frac{\sigma_r}{2} = \frac{\sigma_r}{2} > 0. \tag{6.119}$$

Therefore, $\tilde{\boldsymbol{\Sigma}}_1 = \operatorname{diag}(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_r)$ is invertible. We can now denote the pseudo inverse of $\tilde{\boldsymbol{X}}_1$ by $\tilde{\boldsymbol{X}}_1^\dagger = \tilde{\boldsymbol{U}}_1 \tilde{\boldsymbol{\Sigma}}_1^{-1} \tilde{\boldsymbol{V}}_1^T$. We have

$$\begin{aligned}
\boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{P}_{\tilde{\boldsymbol{U}}_1} &= \boldsymbol{P}_{\boldsymbol{U}_2} \tilde{\boldsymbol{X}}_1 (\tilde{\boldsymbol{X}}_1^\dagger)^T && (\text{since } \boldsymbol{P}_{\tilde{\boldsymbol{U}}_1} = \tilde{\boldsymbol{X}}_1 (\tilde{\boldsymbol{X}}_1^\dagger)^T) \\
&= \boldsymbol{P}_{\boldsymbol{U}_2} (\tilde{\boldsymbol{X}} - \tilde{\boldsymbol{X}}_2)(\tilde{\boldsymbol{X}}_1^\dagger)^T \\
&= \boldsymbol{P}_{\boldsymbol{U}_2} \tilde{\boldsymbol{X}} (\tilde{\boldsymbol{X}}_1^\dagger)^T && (\text{since } \tilde{\boldsymbol{X}}_2 (\tilde{\boldsymbol{X}}_1^\dagger)^T = \boldsymbol{0}) \\
&= \boldsymbol{P}_{\boldsymbol{U}_2} (\boldsymbol{X} + \boldsymbol{\Delta})(\tilde{\boldsymbol{X}}_1^\dagger)^T \\
&= \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} (\tilde{\boldsymbol{X}}_1^\dagger)^T. && (\text{since } \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{X} = \boldsymbol{0})
\end{aligned} \tag{6.120}$$

On the other hand, applying Lemmas 6.7 and 6.5, and the fact that $\left\|\boldsymbol{X}^\dagger\right\|_2 = 1/\sigma_r$, we obtain

$$\left\| \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} (\tilde{\boldsymbol{X}}_1^\dagger)^T \right\|_F \leq \frac{1}{\tilde{\sigma}_r} \|\boldsymbol{\Delta}\|_F . \tag{6.121}$$

From (6.119), we can bound $\tilde{\sigma}_r$ by:

$$\tilde{\sigma}_r > \sigma_r - \frac{\sigma_r}{2} = \frac{\sigma_r}{2}. \tag{6.122}$$

From (6.120), (6.121), and (6.122), we obtain

$$\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}\right\|_F = \left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\tilde{\boldsymbol{X}}_1^{\dagger})^T\right\|_F \leq \frac{1}{\tilde{\sigma}_r}\left\|\boldsymbol{\Delta}\right\|_F < \frac{2}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F. \tag{6.123}$$

Finally, substituting (6.123) back into (6.118) immediately yields $\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}\boldsymbol{\Delta}\right\|_F < \frac{2}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F^2$.

Since in both cases $\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{P}_{\tilde{\boldsymbol{U}}_1}\boldsymbol{\Delta}\right\|_F \leq \frac{2}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F^2$, we conclude from (6.115) that $\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\right\|_F \leq \frac{2}{\sigma_r}\left\|\boldsymbol{\Delta}\right\|_F^2$ for any $\boldsymbol{\Delta}$.

6.8.5.2 Proof of the bound in (6.29)

Taking Frobenius norm on both sides of equation (6.89) and using its absolute homogeneity property, we obtain:

$$\left\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\right\|_F = \left\|\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} + \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2}\right\|_F. \tag{6.124}$$

Applying the triangle inequality to the RHS of (6.124), we have

$$\left\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\right\|_F \leq \left\|\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{X}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\right\|_F + \left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\right\|_F + \left\|\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2}\right\|_F. \tag{6.125}$$

To bound the RHS of (6.125), we proceed by bounding each of the terms on the RHS. The first term on the RHS of (6.125) can be bounded as follows. From (6.107), we have $\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}} \boldsymbol{X} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} = \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2} \boldsymbol{X} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}$. Using Lemmas 6.7 and 6.11, it follows that

$$\begin{aligned} \left\| \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}} \boldsymbol{X} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} \right\|_F &= \left\| \boldsymbol{P}_{\tilde{\boldsymbol{U}}_2} \boldsymbol{X} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} \right\|_F \\ &\leq \left\| \boldsymbol{X} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} \right\|_F \\ &\leq 2 \left\| \boldsymbol{\Delta} \right\|_F . \end{aligned} \tag{6.126}$$

Next, the second term on the RHS of (6.125) can be rewritten as the sum of two orthogonal components

$$\boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} = \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} \boldsymbol{P}_{\boldsymbol{V}_1} + \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} \boldsymbol{P}_{\boldsymbol{V}_2}.$$

By Lemma 6.6, we have

$$\left\| \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} \right\|_F = \sqrt{\left\| \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} \boldsymbol{P}_{\boldsymbol{V}_1} \right\|_F^2 + \left\| \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} \boldsymbol{P}_{\boldsymbol{V}_2} \right\|_F^2}. \tag{6.127}$$

On the one hand, we consider the first term on the RHS of (6.127). Since

$$\begin{aligned} \boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}} \boldsymbol{P}_{\boldsymbol{V}_1} &= (\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2} - \boldsymbol{P}_{\boldsymbol{V}_2}) \boldsymbol{P}_{\boldsymbol{V}_1} \\ &= \boldsymbol{P}_{\tilde{\boldsymbol{V}}_2} \boldsymbol{P}_{\boldsymbol{V}_1}, \qquad \text{(by Lemma 6.3)} \end{aligned}$$

we obtain

$$\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_1}\right\|_F = \left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2}\boldsymbol{P}_{\boldsymbol{V}_1}\right\|_F . \tag{6.128}$$

Applying Lemma 6.7 to the RHS of (6.128) in order to eliminate the three projection matrices, we obtain

$$\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_1}\right\|_F \leq \|\boldsymbol{\Delta}\|_F . \tag{6.129}$$

Similarly, we have

$$\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\boldsymbol{P}_{\boldsymbol{V}_2}\right\|_F \leq \|\boldsymbol{\Delta}\|_F . \tag{6.130}$$

Substituting (6.129), and (6.130) back into (6.127), we have

$$\left\|\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{V}_2}}\right\|_F \leq \sqrt{2}\,\|\boldsymbol{\Delta}\|_F . \tag{6.131}$$

Similarly, we also obtain

$$\left\|\boldsymbol{\delta}_{\boldsymbol{P}_{\boldsymbol{U}_2}}\boldsymbol{\Delta}\boldsymbol{P}_{\tilde{\boldsymbol{V}}_2}\right\|_F \leq \sqrt{2}\,\|\boldsymbol{\Delta}\|_F . \tag{6.132}$$

Substituting (6.126), (6.131), and (6.132) back into (6.125), we obtain

$$\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F \leq 2(1+\sqrt{2})\,\|\boldsymbol{\Delta}\|_F . \tag{6.133}$$

The proof of (6.29) is concluded by taking the minimum between the bounds in (6.133) and (6.28).

6.8.6 Proof of Theorem 6.4

Let us denote

$$\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta}) = \boldsymbol{X}^{\dagger}\boldsymbol{\Delta}^{T}\boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}\boldsymbol{\Delta}^{T}\boldsymbol{X}^{\dagger} + \boldsymbol{P}_{\boldsymbol{U}_2}\boldsymbol{\Delta}(\boldsymbol{X}^{\dagger})^{T}\boldsymbol{\Delta}\boldsymbol{P}_{\boldsymbol{V}_2}.$$

It is straightforward to verify from (6.26) that $\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) = \boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta}) + \boldsymbol{\mathcal{O}}(\|\boldsymbol{\Delta}\|_F^3)$. Thus,

$$\lim_{\epsilon\to 0^+} \sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \frac{\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) - \boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2} = 0. \tag{6.134}$$

Lemma 6.13. *Let f and g be some bounded real-valued functions defined on the set $\mathcal{C}$. Then it holds that*

$$\left|\sup_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x}) - \sup_{\boldsymbol{x}\in\mathcal{C}} g(\boldsymbol{x})\right| \leq \sup_{\boldsymbol{x}\in\mathcal{C}} |f(\boldsymbol{x}) - g(\boldsymbol{x})|\,.$$

Applying Lemma 6.13 to (6.134), we obtain

$$\left|\sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \frac{\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2} - \sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \frac{\|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2}\right| \leq \sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \left|\frac{\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F - \|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2}\right|. \tag{6.135}$$

On the other hand, by the triangle inequality, we have

$$\left|\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F - \|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F\right| \leq \|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) - \boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F \,. \tag{6.136}$$

From (6.135) and (6.136), it holds that

$$\left| \sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \frac{\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2} - \sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \frac{\|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2} \right| \leq \sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \left| \frac{\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta}) - \boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2} \right|. \tag{6.137}$$

Thus, taking both sides of (6.137) to the limit $\epsilon \to 0$ and rearranging terms yield

$$\lim_{\epsilon\to 0^+} \sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \frac{\|\boldsymbol{R}_{\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2} = \lim_{\epsilon\to 0^+} \sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \frac{\|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2}.$$

It now is sufficient to show that

$$\lim_{\epsilon\to 0^+} \sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \frac{\|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F}{\|\boldsymbol{\Delta}\|_F^2} = \frac{1}{\sigma_r\sqrt{3}}. \tag{6.138}$$

Indeed, due to the orthogonality among the addends, we have

$$\begin{aligned} \|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F^2 &= \left\| \boldsymbol{X}^\dagger \boldsymbol{\Delta}^T \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} + \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} \boldsymbol{\Delta}^T \boldsymbol{X}^\dagger + \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} (\boldsymbol{X}^\dagger)^T \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} \right\|_F^2 \\ &= \left\| \boldsymbol{X}^\dagger \boldsymbol{\Delta}^T \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} \right\|_F^2 + \left\| \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} \boldsymbol{\Delta}^T \boldsymbol{X}^\dagger \right\|_F^2 + \left\| \boldsymbol{P}_{\boldsymbol{U}_2} \boldsymbol{\Delta} (\boldsymbol{X}^\dagger)^T \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_2} \right\|_F^2 . \end{aligned} \tag{6.139}$$

Using the definition of $\boldsymbol{E}$ in (6.4), (6.139) can be represented as

$$\begin{aligned}\|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F^2 &= \left\|\boldsymbol{U}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{E}_{21}^T\boldsymbol{E}_{22}\boldsymbol{V}_2^T\right\|_F^2 + \left\|\boldsymbol{U}_2\boldsymbol{E}_{22}\boldsymbol{E}_{12}^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{V}_1^T\right\|_F^2 + \left\|\boldsymbol{U}_2\boldsymbol{E}_{21}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{E}_{12}\boldsymbol{V}_2^T\right\|_F^2 \\ &= \left\|\boldsymbol{\Sigma}_1^{-1}\boldsymbol{E}_{21}^T\boldsymbol{E}_{22}\right\|_F^2 + \left\|\boldsymbol{E}_{22}\boldsymbol{E}_{12}^T\boldsymbol{\Sigma}_1^{-1}\right\|_F^2 + \left\|\boldsymbol{E}_{21}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{E}_{12}\right\|_F^2, \qquad (6.140)\end{aligned}$$

where the second equality stems from Lemma 6.7. Using Lemma 6.5 and the fact that $\left\|\boldsymbol{\Sigma}_1^{-1}\right\|_2 = 1/\sigma_r$, we can bound the RHS of (6.140) by

$$\|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F^2 \leq \frac{1}{\sigma_r^2}\Big(\|\boldsymbol{E}_{21}\|_F^2\,\|\boldsymbol{E}_{22}\|_F^2 + \|\boldsymbol{E}_{22}\|_F^2\,\|\boldsymbol{E}_{12}\|_F^2 + \|\boldsymbol{E}_{12}\|_F^2\,\|\boldsymbol{E}_{21}\|_F^2\Big). \qquad (6.141)$$

Lemma 6.14 (Chebyshev's sum inequality [93])**.** *For any $a, b, c \in \mathbb{R}$, we have*

$$3(ab + bc + ca) \leq (a + b + c)^2.$$

Applying Lemma 6.14 to (6.141) with $a = \|\boldsymbol{E}_{21}\|_F^2, b = \|\boldsymbol{E}_{22}\|_F^2$ and $c = \|\boldsymbol{E}_{12}\|_F^2$, we obtain

$$\begin{aligned}\|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F^2 &\leq \frac{1}{\sigma_r^2}\frac{\left(\|\boldsymbol{E}_{21}\|_F^2 + \|\boldsymbol{E}_{22}\|_F^2 + \|\boldsymbol{E}_{12}\|_F^2\right)^2}{3} \\ &\leq \frac{\left(\|\boldsymbol{E}_{11}\|_F^2 + \|\boldsymbol{E}_{12}\|_F^2 + \|\boldsymbol{E}_{21}\|_F^2 + \|\boldsymbol{E}_{22}\|_F^2\right)^2}{3\sigma_r^2} = \frac{\|\boldsymbol{E}\|_F^4}{3\sigma_r^2} = \frac{\|\boldsymbol{\Delta}\|_F^4}{3\sigma_r^2}, \qquad (6.142)\end{aligned}$$

where the last equation stems from $\|\boldsymbol{E}\|_F = \left\|\boldsymbol{U}^T\boldsymbol{\Delta}\boldsymbol{V}\right\|_F = \|\boldsymbol{\Delta}\|_F$. From (6.142),

taking the square root and then taking the supremum yield

$$\sup_{\|\boldsymbol{\Delta}\|_F=\epsilon} \|\boldsymbol{R}_{2\boldsymbol{X}}(\boldsymbol{\Delta})\|_F \leq \frac{\|\boldsymbol{\Delta}\|_F^2}{\sigma_r\sqrt{3}}. \tag{6.143}$$

To show that (6.143) implies (6.138), we describe a particular choice of $\boldsymbol{\Delta}$ such that the inequality holds. Let us choose

$$\boldsymbol{\Delta}(\epsilon) \triangleq \frac{\epsilon}{\sqrt{3}}\big(\boldsymbol{u}_r\boldsymbol{v}_{r+1}^T + \boldsymbol{u}_{r+1}\boldsymbol{v}_r^T + \boldsymbol{u}_{r+1}\boldsymbol{v}_{r+1}^T\big),$$

where $\boldsymbol{u}_r, \boldsymbol{u}_{r+1}, \boldsymbol{v}_r$, and $\boldsymbol{v}_{r+1}$ are the corresponding left and right singular vectors of $\boldsymbol{X}$. Similar to (6.101), one can verify that $\|\boldsymbol{\Delta}(\epsilon)\|_F = \epsilon$. In addition, from Proposition 6.2, we have

$$\begin{aligned}
&\boldsymbol{E}_{12} = \boldsymbol{U}_1^T\boldsymbol{\Delta}(\epsilon)\boldsymbol{V}_2 = \frac{\epsilon}{\sqrt{3}}\boldsymbol{e}_r^r(\boldsymbol{e}_1^{n-r})^T,\ \boldsymbol{E}_{21} = \boldsymbol{U}_2^T\boldsymbol{\Delta}(\epsilon)\boldsymbol{V}_1 = \frac{\epsilon}{\sqrt{3}}\boldsymbol{e}_1^{m-r}(\boldsymbol{e}_r^r)^T,\\
&\boldsymbol{E}_{22} = \boldsymbol{U}_2^T\boldsymbol{\Delta}(\epsilon)\boldsymbol{V}_2 = \frac{\epsilon}{\sqrt{3}}\boldsymbol{e}_1^{m-r}(\boldsymbol{e}_1^{n-r})^T.
\end{aligned} \tag{6.144}$$

Substituting (6.144) back into (6.140) yields

$$\begin{aligned}
&\big\|\boldsymbol{R}_{2\boldsymbol{X}}\big(\boldsymbol{\Delta}(\epsilon)\big)\big\|_F^2\\
&\quad= \left\|\frac{\epsilon^2}{3}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{e}_r^r(\boldsymbol{e}_1^{m-r})^T\boldsymbol{e}_1^{m-r}(\boldsymbol{e}_1^{n-r})^T\right\|_F^2 + \left\|\frac{\epsilon^2}{3}\boldsymbol{e}_1^{m-r}(\boldsymbol{e}_1^{n-r})^T\boldsymbol{e}_1^{n-r}(\boldsymbol{e}_r^r)^T\boldsymbol{\Sigma}_1^{-1}\right\|_F^2\\
&\qquad\qquad + \left\|\frac{\epsilon^2}{3}\boldsymbol{e}_1^{m-r}\boldsymbol{e}_r^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{e}_r(\boldsymbol{e}_1^{n-r})^T\right\|_F^2\\
&\quad= \frac{\epsilon^4}{9}\left(\frac{1}{\sigma_r^2}+\frac{1}{\sigma_r^2}+\frac{1}{\sigma_r^2}\right) = \frac{\|\boldsymbol{\Delta}\|_F^4}{3\sigma_r^2}. \qquad \text{(since } \|\boldsymbol{\Delta}(\epsilon)\|_F = \epsilon\text{)}
\end{aligned}$$

Therefore, the equality in (6.143) holds when $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\epsilon)$, for any $\epsilon > 0$. This completes our proof of the theorem.

6.8.6.1 Proof of Lemma 6.13

Since $f(\boldsymbol{x}) - g(\boldsymbol{x}) \leq |f(\boldsymbol{x}) - g(\boldsymbol{x})|$, we have $f(\boldsymbol{x}) \leq |f(\boldsymbol{x}) - g(\boldsymbol{x})| + g(\boldsymbol{x})$. Taking the supremum yields

$$\begin{aligned}\sup_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x}) &\leq \sup_{\boldsymbol{x}\in\mathcal{C}}\Big\{ |f(\boldsymbol{x}) - g(\boldsymbol{x})| + g(\boldsymbol{x}) \Big\} \\ &\leq \sup_{\boldsymbol{x}\in\mathcal{C}} |f(\boldsymbol{x}) - g(\boldsymbol{x})| + \sup_{\boldsymbol{x}\in\mathcal{C}} g(\boldsymbol{x}).\end{aligned}$$

Thus, we have

$$\sup_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x}) - \sup_{\boldsymbol{x}\in\mathcal{C}} g(\boldsymbol{x}) \leq \sup_{\boldsymbol{x}\in\mathcal{C}} |f(\boldsymbol{x}) - g(\boldsymbol{x})| . \tag{6.145}$$

Changing the roles of f and g, we also obtain

$$\sup_{\boldsymbol{x}\in\mathcal{C}} g(\boldsymbol{x}) - \sup_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x}) \leq \sup_{\boldsymbol{x}\in\mathcal{C}} |f(\boldsymbol{x}) - g(\boldsymbol{x})| . \tag{6.146}$$

Our inequality follows on combining (6.145) and (6.146).

Chapter 7: On Local Convergence of Iterative Hard Thresholding for Matrix Completion[1]

Iterative hard thresholding (IHT) has gained in popularity over the past decades in large-scale optimization. However, convergence property of this method has only been explored recently in non-convex settings. In matrix completion, existing works often focus on the guarantee of global convergence of IHT via standard assumptions such as incoherence property and uniform sampling. While such analysis provides a global upper bound on the linear convergence rate, it does not describe the actual performance of IHT in practice. In this chapter, we provide a novel insight into the local convergence of a specific variant of IHT for matrix completion. We uncover the exact linear rate of IHT in a closed-form expression and identify the region of convergence in which the algorithm is guaranteed to converge. Furthermore, we utilize random matrix theory to study the linear rate of convergence of IHTSVD for large-scale matrix completion. We find that asymptotically, the rate can be expressed in closed form in terms of the relative rank and the sampling rate. Finally, we present various numerical results to verify the aforementioned theoretical analysis.

[1]This work is currently under review and available at `https://arxiv.org/abs/2112.14733`.

7.1 Introduction

Matrix completion is a fundamental problem that arises in many areas of signal processing and machine learning such as collaborative filtering [176, 188, 189, 200], system identification [136,137,156] and dimension reduction [31,228]. The problem can be explained as follows. Let $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ be the underlying matrix with rank r and Ω be the set of locations corresponding to the observed entries of $\boldsymbol{M}$, i.e., $(i, j) \in \Omega$ if M_{ij} is observed. The goal is to recover the unknown entries of $\boldsymbol{M}$, belonging to the complement set $\overline{\Omega}$.

To understand the feasibility of matrix completion, let us describe $\boldsymbol{M}$ as

$$\boldsymbol{M} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T,$$

where σ_i is the i-th largest singular value of $\boldsymbol{M}$, $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the corresponding left and right singular vectors. Since each set of the left and right singular vectors are orthonormal, the degrees of freedom of matrix completion is given by

$$r + \sum_{i=1}^{r} (n_1 - i) + \sum_{j=1}^{r} (n_2 - j) = (n_1 + n_2 - r)r,$$

which is significantly less than the total number of entries in $\boldsymbol{M}$ when r is small. This implies possibility of recovering the entire matrix even when only a few entries are observed. However, not every matrix with more than $(n_1 + n_2 - r)r$ observed entries can be completed. For instance, if an entire column (or row) of a rank-one matrix is missing, then the matrix cannot be recovered. Similarly, if a

low-rank matrix contains too many zero entries, then the observed entries might end up being all zero, thereby not providing any clue about the missing entries. The aforementioned argument motivates the two standard assumptions in matrix completion: the incoherence condition and the random sampling model. Under these assumptions, Candès and Recht [33] showed that matrix completion can be solved exactly for most settings of the low-rank matrix $\boldsymbol{M}$ and the sampling set Ω. This breakthrough has started a long line of research on efficient methods for solving matrix completion.

In the same work, Candès and Recht [33] proposed a convex relaxation approach to matrix completion, replacing the original linearly constrained rank minimization problem by a linearly constrained nuclear norm minimization problem. Their result leads to a well-known class of proximal-type algorithms for nuclear norm minimization [27, 107, 145, 205] with rigorous mathematical guarantees and extensions of classic acceleration techniques. Nonetheless, convex-relaxed methods are generally considered slow compared to their non-convex counterparts in practice. While interior-point methods for solving the nuclear norm minimization problem is computationally expensive and even infeasible for large matrices, proximal-type algorithms suffer from slow convergence due to the conservative nature of the soft-thresholding operator [117, 213].

Another approach to matrix completion is known as iterative hard thresholding. To address the computational concern from the use of convex relaxation, IHT methods have been proposed to directly solve the non-convex rank minimization problem [79, 104]. Each IHT iteration takes one step in the opposite direction of

the gradient and another step projecting the result onto the set of rank-r matrices. Since the process resembles hard-thresholding singular values, we refer to the class of algorithms using this technique as iterative hard thresholding. When the solution is low-rank, hard-thresholding algorithms is more efficient than their soft-thresholding counterparts in both computational complexity per iteration and convergence speed. Variants of plain IHT with faster convergence have also been developed, including normalized IHT [201], conjugate gradient IHT [18], Nesterov's accelerated gradient IHT [213], Heavy-Ball IHT [214], just to name a few. The drawback of IHT methods, however, is the lack of mathematical guarantees on their convergence behavior. As pointed out in [104], the restricted isometry property (RIP), which is widely-used in establishing the global convergence in matrix sensing, does not hold for matrix completion. Therefore, the global convergence of IHT methods for matrix completion is still an open question. Until recently, the only guarantee on the global convergence of a IHT method, to the best of our knowledge, is provided in [105]. In their work, the authors considered a variant of the singular value projection (SVP) algorithm with resampling scheme and proved the fast linear convergence of the proposed algorithm with a sample complexity that depends on the condition number and desired accuracy. Notwithstanding, this result imposes some limitations at conceptual, practical and theoretical levels due to the requirement of resampling [199]. In a different perspective, local convergence of IHT methods has also been studied by Chunikhina *et. al.* [46]. In particular, by considering a special case of the SVP algorithm with unit step size, called iterative hard-thresholded singular value decomposition (IHTSVD), the au-

thors showed that IHTSVD converges linearly to the solution $\boldsymbol{M}$ as long as the algorithm is initialized close enough to $\boldsymbol{M}$. Consequently, this analysis explains the superior performance of IHT methods over proximal-type methods in practice.[2] A similar approach can be found in the unpublished work of Lai and Varghese [120]. However, we remark that while the later work proves the existence of an upper bound on the linear convergence rate of IHTSVD, the former provides an exact expression of the rate that depends directly on the structure of $\boldsymbol{M}$ and Ω.

The most popular approach to matrix completion is non-convex factorization. This approach stems from the Burer-Monteiro factorization [26], whereby the low-rank matrix is viewed as a product of two low-rank components. The resulting least squares problem is unconstrained albeit non-convex. Recent progress in this approach has shown that any local minimum of the re-parameterized problem is also a global minimum [76,199]. Thus, basic optimization procedures such as gradient descent [43,144,199] and alternating minimization [40,91,92,106] can provably find the global solution at a linear convergence rate. The exact linear convergence rate of gradient descent for matrix completion has recently been studied by Vu and Raich [215]. In Table 7.1, we summarize the aforementioned approaches to matrix completion and the corresponding algorithms existing in the literature.

This chapter is developed based on the work of Chunikhina *et. al.* [46] on the local convergence of the IHTSVD algorithm for matrix completion. Our main contribution is three-fold. First, we propose a novel analysis of the local convergence

[2]Convergence guarantees on proximal-type methods for matrix completion are often sublinear [27, 205].

Table 7.1: Three common formulations of matrix completion problem.

Problem formulation	**Description**	**Algorithms**
Linearly constrained nuclear norm minimization	$\min_{\boldsymbol{X}\in\mathbb{R}^{n_1\times n_2}} \|\boldsymbol{X}\|_* \text{ s.t. } X_{ij}=M_{ij}, \quad (i,j)\in\Omega$	Semi-definite programming (SDP) [33], singular value thresholding (SVT) [27], accelerated proximal gradient (APG) [205], conditional gradient descent (CGD) [23, 102, 172]
Rank-constrained least squares	$\min_{\boldsymbol{X}\in\mathbb{R}^{n_1\times n_2}} \sum_{(i,j)\in\Omega}(X_{ij}-M_{ij})^2 \text{ s.t. } \operatorname{rank}(\boldsymbol{X})\leq r$	Singular value projection (SVP) [104], normalized IHT (NIHT) [201], conjugate gradient IHT (CGIHT) [18], iterative hard-thresholded SVD (IHTSVD) [46], accelerated IHT [213, 214]
Low-rank factorization	$\min_{\boldsymbol{Y}\in\mathbb{R}^{n_1\times r},\boldsymbol{Z}\in\mathbb{R}^{n_2\times r}} \sum_{(i,j)\in\Omega}((\boldsymbol{Y}\boldsymbol{Z}^\top)_{ij}-M_{ij})^2$	Alternating minimization (AM) [91, 106], gradient descent (GD) [144, 199], projected gradient descent (PGD) [26, 43], stochastic gradient descent (SGD) [199]

of IHTSVD for matrix completion. The proposed analysis establishes the region of convergence that is proportional to the least non-zero singular value of $\boldsymbol{M}$. Moreover, we show that the convergence is asymptotically linear and the exact rate can be described in a closed-form expression of the projections onto the (left and right) null spaces of $\boldsymbol{M}$ and the sampling pattern Ω. Second, based on the exact linear rate, we utilize random matrix theory to study the asymptotic behavior of IHTSVD in large-scale matrix completion. As the size of $\boldsymbol{M}$ grows to infinity, we uncover the linear rate of IHTSVD converges to a deterministic constant that can be expressed in closed form in terms of the relative rank and the sampling rate. Finally, we present extensive results to verify our proposed exact rate of convergence as well as the asymptotic rate of IHTSVD in large-scale settings.

7.2 Preliminaries

7.2.1 Notation

Throughout the chapter, we use the notations $\|\cdot\|_F$, $\|\cdot\|_2$, and $\|\cdot\|_{2,\infty}$ to denote the Frobenius norm, the spectral norm and the l_2/l_∞ norm (i.e., the largest l_2 norm of the rows) of a matrix, respectively. Occasionally, $\|\cdot\|_2$ is used on a vector to denote the Euclidean norm. The notation $[n]$ refers to the set $\{1, 2, \ldots, n\}$. Boldfaced symbols are reserved for vectors and matrices. In addition, let $\boldsymbol{I}_n$ denote the $n \times n$ identity matrix. $\otimes$ denotes the Kronecker product between two matrices.

For a matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, X_{ij} refers to the (i, j) element of $\boldsymbol{X}$. We denote

$\sigma_{\max}(\boldsymbol{X})$ and $\sigma_{\min}(\boldsymbol{X})$ as the largest and smallest singular values of $\boldsymbol{X}$, respectively, and denote $\kappa(\boldsymbol{X}) = \sigma_{\max}(\boldsymbol{X})/\sigma_{\min}(\boldsymbol{X})$ as the condition number of $\boldsymbol{X}$. $\mathrm{vec}(\boldsymbol{X})$ denotes the vectorization of $\boldsymbol{X}$ by stacking its columns on top of one another. Let $\boldsymbol{F}(\boldsymbol{X})$ be a matrix-valued function of $\boldsymbol{X}$. Then, for some $k > 0$, we use $\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{\mathcal{O}}(\|\boldsymbol{X}\|_F^k)$ to imply

$$\lim_{\delta \to 0} \sup_{\|\boldsymbol{X}\|_F = \delta} \frac{\|\boldsymbol{F}(\boldsymbol{X})\|_F}{\|\boldsymbol{X}\|_F^k} < \infty.$$

7.2.2 Background

Let us use $\boldsymbol{M}$ to denote the underlying $n_1 \times n_2$ real matrix with rank

$$1 \leq r \leq m = \min\{n_1, n_2\}. \tag{7.1}$$

The sampling set Ω is a subset of the Cartesian product $[n_1] \times [n_2]$, with cardinality of $1 \leq s < n_1 n_2$. Furthermore, the orthogonal projection associated with Ω is given in the following:

Definition 7.1. *The orthogonal projection onto the set of matrices supported in* Ω *is defined as a linear operator* $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ *satisfying*

$$[\mathcal{P}_\Omega(\boldsymbol{X})]_{ij} = \begin{cases} X_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{if } (i,j) \in \overline{\Omega}, \end{cases}$$

where $\overline{\Omega}$ *denotes the complement set of* Ω.

If we consider vector spaces instead of matrix spaces, the orthogonal projection $\mathcal{P}_\Omega$ can also be viewed as a selection matrix corresponding to Ω:

Definition 7.2. *The selection matrix $\boldsymbol{S}_\Omega \in \mathbb{R}^{n_1 n_2 \times s}$ comprises a subset of s columns of the identity matrix of dimension $n_1 n_2$ such that*

$$\begin{cases} \boldsymbol{S}_\Omega^\top \boldsymbol{S}_\Omega = \boldsymbol{I}_s, \\ \operatorname{vec}\big(\mathcal{P}_\Omega(\boldsymbol{X})\big) = \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \operatorname{vec}(\boldsymbol{X}). \end{cases}$$

Corresponding to the complement set $\overline{\Omega}$, we also define similar notations for $\mathcal{P}_{\overline{\Omega}} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ and $\boldsymbol{S}_{\bar{\Omega}} \in \mathbb{R}^{n_1 n_2 \times (n_1 n_2 - s)}$.

Next, using the notation of $\mathcal{P}_\Omega$, we can formulate the matrix completion problem as follows:

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2} \left\| \mathcal{P}_\Omega(\boldsymbol{X} - \boldsymbol{M}) \right\|_F^2 \text{ s.t. } \operatorname{rank}(\boldsymbol{X}) \leq r. \tag{7.2}$$

One natural approach to the optimization problem (7.2) is projected gradient descent. Starting at some $\boldsymbol{X}^{(0)}$, we iteratively update the current matrix by (i) taking a step in the opposite direction of the gradient and (ii) projecting the result back onto the set of matrices with rank less than or equal to r. It follows that

$$\boldsymbol{X}^{(k+1)} = \mathcal{P}_r\big(\boldsymbol{X}^{(k)} - \eta \mathcal{P}_\Omega(\boldsymbol{X}^{(k)} - \boldsymbol{M})\big), \tag{7.3}$$

where η is the step size and $\mathcal{P}_r$ is the rank-r projection (formally defined later in Definition 7.3). Due to the singular value truncating nature of the projection

$\mathcal{P}_r$, PGD is often referred as the iterative hard thresholding (IHT) method for matrix completion [79]. In [104], IHT with step size $\eta = n_1 n_2/s$ is also named as the Singular Value Projection (SVP) algorithm for matrix completion. In the literature, PGD with step size $\eta = n_1 n_2/s$ is also known as the Singular Value Projection (SVP) algorithm for matrix completion [104]. It is interesting to note that under certain assumptions, [105] showed that the algorithm enjoys a fast global linear convergence with this choice of step size. On the other hand, setting the step size $\eta = 1$ yields the following update

$$\begin{aligned} \boldsymbol{X}^{(k+1)} &= \mathcal{P}_r\big(\boldsymbol{X}^{(k)} - \mathcal{P}_\Omega(\boldsymbol{X}^{(k)} - \boldsymbol{M})\big) \\ &= \mathcal{P}_r\big(\mathcal{P}_{\overline{\Omega}}(\boldsymbol{X}^{(k)}) + \mathcal{P}_\Omega(\boldsymbol{M})\big). \end{aligned} \tag{7.4}$$

Note that $\mathcal{P}_{\bar{\Omega}}(\boldsymbol{X}^{(k)}) + \mathcal{P}_\Omega(\boldsymbol{M})$ is a linear orthogonal projection of $\boldsymbol{X}^{(k)}$ onto the set of matrices with the same support as $\boldsymbol{M}$ in Ω. This motivates the IHTSVD algorithm [46] that alternates between two projection steps: the set of low-rank matrices and the projection onto the set of matrices with the same support as $\boldsymbol{M}$ in Ω (see Algorithm 7.1). This chapter, developed based on [46], focuses on local convergence properties of IHTSVD. Compared to the existing global convergence analysis for matrix completion, our setting does not require certain assumptions such as the incoherence of $\boldsymbol{M}$, the uniform randomness of Ω, and the low sample complexity, e.g., $s = \mathcal{O}(r^5 n \log n)$ in [105]. We also note that the proposed analysis can be extended to other variants of PGD with different step sizes.

Finally, we present a formal definition of the rank-r projection. Consider matrix

$\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ with the singular value decomposition

$$\boldsymbol{X} = \sum_{i=1}^{m} \sigma_i(\boldsymbol{X}) \boldsymbol{u}_i(\boldsymbol{X}) \boldsymbol{v}_i^\top(\boldsymbol{X}),$$

where $\sigma_1(\boldsymbol{X}) \geq \ldots \geq \sigma_m(\boldsymbol{X}) \geq 0$ are the singular values of $\boldsymbol{X}$ and $\{\boldsymbol{u}_1(\boldsymbol{X}), \ldots, \boldsymbol{u}_m(\boldsymbol{X})\}$, $\{\boldsymbol{v}_1(\boldsymbol{X}), \ldots, \boldsymbol{v}_m(\boldsymbol{X})\}$ are the sets of left and right singular vectors of $\boldsymbol{X}$, respectively.

Definition 7.3. *The rank-r projection of $\boldsymbol{X}$ is defined as*

$$\mathcal{P}_r(\boldsymbol{X}) = \sum_{i=1}^{r} \sigma_i(\boldsymbol{X}) \boldsymbol{u}_i(\boldsymbol{X}) \boldsymbol{v}_i^\top(\boldsymbol{X}).$$

The rank-r projection of $\boldsymbol{X}$ is unique if and only if $\sigma_r(\boldsymbol{X}) > \sigma_{r+1}(\boldsymbol{X})$ or $\sigma_r(\boldsymbol{X}) = 0$ [61]. Since $\mathcal{P}_r(\boldsymbol{X})$ zeroes out all the small singular value of $\boldsymbol{X}$, it is often referred as the singular value hard-thresholding operator. Since $\boldsymbol{M}$ is a rank-r matrix, we have

$$\boldsymbol{M} = \mathcal{P}_r(\boldsymbol{M}) = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^\top,$$

where $\Sigma_r = \text{diag}(\sigma_1, \ldots, \sigma_r)$ contains the singular values of $\boldsymbol{M}$ and $\boldsymbol{U}_r = [u_1, \ldots, u_r] \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{V}_r = [v_1, \ldots, v_r] \in \mathbb{R}^{n_2 \times r}$ are comprised of the first r left and right singular vectors of $\boldsymbol{M}$, respectively.[3] Denote $\boldsymbol{U}_\perp = [u_{r+1}, \ldots, u_{n_1}] \in \mathbb{R}^{n_1 \times (n_1 - r)}$ and $\boldsymbol{V}_\perp = [v_{r+1}, \ldots, v_{n_2}] \in \mathbb{R}^{n_2 \times (n_2 - r)}$. The projections onto the left and right null

[3] In the rest of this chapter, we omit the parameter in the notation of the singular values and the singular vectors of $\boldsymbol{M}$ for simplicity.

spaces of $\boldsymbol{M}$ are uniquely defined as

$$\boldsymbol{P}_{\boldsymbol{U}_\perp} = \boldsymbol{U}_\perp \boldsymbol{U}_\perp^\top = \boldsymbol{I}_{n_1} - \sum_{i=1}^{r} u_i u_i^\top,$$
$$\boldsymbol{P}_{\boldsymbol{V}_\perp} = \boldsymbol{V}_\perp \boldsymbol{V}_\perp^\top = \boldsymbol{I}_{n_2} - \sum_{i=1}^{r} v_i v_i^\top.$$

7.2.3 Related Work

Traditional approaches to matrix completion often make assumptions on the incoherence of the underlying matrix $\boldsymbol{M}$ and the randomness of the sampling set. First, the incoherence condition for matrix completion, introduced by Candès and Recht [33], is stated as:

Assumption 7.1 (Incoherence)**.** *The matrix $\boldsymbol{M} = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^\top$ is μ-incoherent, i.e.,*

$$\|\boldsymbol{U}_r\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n_1}} \text{ and } \|\boldsymbol{V}_r\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n_2}}.$$

Intuitively, an incoherent matrix has well-spread singular vectors and is less likely in the null space of the sampling operator. A common setting that generates incoherent matrices is the random orthogonal model:

Definition 7.4 (Random orthogonal model)**.** *The Haar measure provides a uniform and translation-invariant distribution over the group of $n \times n$ orthogonal matrices $\mathbb{O}(n)$. $\boldsymbol{M}$ is said to follow a random orthogonal model if $\boldsymbol{U}_r$ and $\boldsymbol{V}_r$ are sub-matrices of Haar-distributed matrices in $\mathbb{O}(n_1)$ and $\mathbb{O}(n_2)$, respectively.*

Second, in order to avoid adversarial patterns in the sampling set, it is also common to assume that each entry in Ω is selected randomly:

Assumption 7.2 (Uniform sampling)**.** *The sampling set Ω is obtained by selecting s elements uniformly at random from the Cartesian product $[n_1] \times [n_2]$.*

We note that a similar but not equivalent assumption on the sampling set is the Bernoulli model in which each entry of $\boldsymbol{M}$ is observed independently with probability s/n_1n_2 [199]. Under these two standard assumptions, Candès and Recht [33] showed that symmetric matrix completion of size n can be solved exactly provided that the number of observations is sufficiently large, i.e., $s = \mathcal{O}(n^{1.2}r \log n)$. Later on, global convergence guarantees for various matrix-completion algorithms have been actively developed, with improved bounds on the sample complexity. Examples of these works include [92, 105, 144, 174, 199]. It is worthwhile mentioning that ideally, one would like to recover the low-rank matrix from a minimum number of observations, which is in the order the degrees of freedom of the problem, i.e., $\mathcal{O}(nr)$.

In this chapter, we study the convergence of IHT for matrix completion from a different perspective. Without any assumptions about the incoherence of $\boldsymbol{M}$ and the randomness of the sampling set Ω, we identify a deterministic condition on the structure of $\boldsymbol{M}$ and Ω such that the local linear convergence of IHTSVD can be guaranteed. Compared to the aforementioned bounds on the global convergence rate, our result is exact and tighter thanks to the exploitation of the local structure of the problem. Our technique utilizes the recently developed error bound for

the first-order Taylor expansion of the rank-r projection, proposed by Vu *et. al.* in [211]. The result is rephrased below.

Proposition 7.1 (Rephrased from [211])**.** *For any* $\boldsymbol{\Delta} \in \mathbb{R}^{n_1 \times n_2}$*, we have*

$$\mathcal{P}_r(\boldsymbol{M} + \boldsymbol{\Delta}) = \boldsymbol{M} + \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{\Delta}), \tag{7.5}$$

where the residual $\boldsymbol{R} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ *satisfies:*

$$\|\boldsymbol{R}(\boldsymbol{\Delta})\|_F \leq \frac{c_1}{\sigma_r} \|\boldsymbol{\Delta}\|_F^2 ,$$

for some universal constant $1 + 1/\sqrt{2} \leq c_1 \leq 4(1 + \sqrt{2})$*.*

The rest of the chapter is organized as follows. In Section 7.3, we provide the local convergence analysis of IHTSVD for matrix completion and the proof of the main result. Next, Section 7.4 presents a brief summary of related results in random matrix theory, followed by our novel result on the asymptotic behavior of the convergence rate in large-scale settings. The numerical results to verify the analysis in Sections 7.3 and 7.4 are given in Section 7.5. Finally, we put the detailed proofs of all the main theorems and lemmas in the appendix.

7.3 Local Convergence of IHTSVD

This section presents our analysis of local convergence of IHTSVD. First, we leverage the results in perturbation analysis to identify the Taylor series expansion of

Algorithm 7.1 IHTSVD

Require: $\mathcal{P}_\Omega(\boldsymbol{M}), r, K$
Ensure: $\boldsymbol{X}^{(K)}$
1: Initialize $\boldsymbol{X}^{(0)}$
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: $\quad \boldsymbol{Y}^{(k)} = \mathcal{P}_r\big(\boldsymbol{X}^{(k)}\big)$
4: $\quad \boldsymbol{X}^{(k+1)} = \mathcal{P}_{\overline{\Omega}}\big(\boldsymbol{Y}^{(k)}\big) + \mathcal{P}_\Omega(\boldsymbol{M})$

the rank-r projection. Next, the approximation allows us to derive the nonlinear difference equation that describes the change in the distance to the local optimum through IHT iterations. Closed-form expressions of the asymptotic convergence rate and the region of convergence are also given as a result of our analysis.

7.3.1 Main Result

Our local convergence result is stated as follows:

Theorem 7.1. *Let* $\{\boldsymbol{X}^{(k)}\}$ *be the sequence of matrices generated by Algorithm 7.1, i.e., for* $k = 0, 1, \ldots$*:*

$$\boldsymbol{X}^{(k+1)} = \mathcal{P}_{\overline{\Omega}}\big(\mathcal{P}_r(\boldsymbol{X}^{(k)})\big) + \mathcal{P}_\Omega(\boldsymbol{M}) \tag{7.6}$$

and assume that $\lambda_{\min}(\boldsymbol{H}) > 0$ *and* $\boldsymbol{X}^{(0)}$ *satisfies:*

$$\big\|\boldsymbol{X}^{(0)} - \boldsymbol{M}\big\|_F < \frac{\lambda_{\min}(\boldsymbol{H})}{c_1}\sigma_r, \tag{7.7}$$

where

$$\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^{\top}(\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})\boldsymbol{S}_{\bar{\Omega}} \quad \in \mathbb{R}^{(n_1 n_2 - s)\times(n_1 n_2 - s)}. \tag{7.8}$$

Then, $\left\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\right\|_F$ *converge asymptotically at a linear rate of*

$$\rho = 1 - \lambda_{\min}(\boldsymbol{H}). \tag{7.9}$$

Specifically, for any $\epsilon > 0$, $\left\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\right\|_F \leq \epsilon \left\|\boldsymbol{X}^{(0)} - \boldsymbol{M}\right\|_F$ *for all* k *such that*

$$k \geq N(\epsilon) = \left\lceil \frac{\log(1/\epsilon) + c_2}{\log(1/(1 - \lambda_{\min}(\boldsymbol{H})))} \right\rceil. \tag{7.10}$$

where $c_2 > 0$ *is a constant depending only on* $\boldsymbol{X}^{(0)}$ *and* $\boldsymbol{M}$.

Theorem 7.1 provides a closed-form expression of the linear convergence rate of IHTSVD for matrix completion. As can be seen in (7.10), the speed of convergence depends strongly on how close the smallest eigenvalue of $\boldsymbol{H}$ is to zero: as $\lambda_{\min}(\boldsymbol{H})$ approaches 0, the number of iterations needed to reach a relative accuracy of ϵ, i.e., $N(\epsilon)$, grows to infinity. In fact, when $\lambda_{\min}(\boldsymbol{H}) = 0$, the condition in (7.7) cannot be satisfied and hence, there is no linear convergence guarantee provided by our theorem in this case. On the other hand, from (7.8), one can verify that all eigenvalues of $\boldsymbol{H}$ lie between 0 and 1 since the norm of either a projection matrix or a selection matrix is less than or equal to 1. This combined with the aforementioned condition that $\lambda_{\min}(\boldsymbol{H}) > 0$ ensures the linear convergence rate ρ in (7.9) belongs to $[0, 1)$.

Remark 7.1. *Theorem 7.1 does not guarantee linear convergence when* $\lambda_{\min}(\boldsymbol{H}) = 0$. *Interestingly, one such situation is when* $\boldsymbol{H}$ *is* ***rank-deficient****. Let us represent*

$$\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^{\top}(\boldsymbol{V}_{\perp} \otimes \boldsymbol{U}_{\perp})(\boldsymbol{V}_{\perp} \otimes \boldsymbol{U}_{\perp})^{\top}\boldsymbol{S}_{\bar{\Omega}} = \boldsymbol{W}\boldsymbol{W}^{\top},$$

where $\boldsymbol{W} = \boldsymbol{S}_{\bar{\Omega}}^{\top}(\boldsymbol{V}_{\perp} \otimes \boldsymbol{U}_{\perp}) \in \mathbb{R}^{(n_1 n_2 - s)\times(n_1 - r)(n_2 - r)}$. *If* $\boldsymbol{W}$ *is a tall matrix, i.e.,*

$$s < (n_1 + n_2 - r)r, \tag{7.11}$$

then it follows that $\boldsymbol{H}$ *is rank-deficient and* $\lambda_{\min}(\boldsymbol{H}) = 0$. *We note that in this case the number of sampled entries is less than the degrees of freedom of the problem.*

Remark 7.2. *When* $s \geq (n_1 + n_2 - r)r$, *it is possible that* $\lambda_{\min}(\boldsymbol{H}) = 0$ *for certain (adversarial) sampling patterns. For example, consider a* 3×2 *rank-*1 *matrix*

$$\boldsymbol{M} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \end{bmatrix}^{\top}.$$

One choice of the matrices $\boldsymbol{U}_{\perp}$ *and* $\boldsymbol{V}_{\perp}$ *is*

$$\boldsymbol{U}_{\perp} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \boldsymbol{V}_{\perp} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

If we observe $s = 4$ *entries of the first two rows of* $\boldsymbol{M}$, *the selection matrix* $\boldsymbol{S}_{\bar{\Omega}}$ *is*

given by

$$\boldsymbol{S}_{\bar{\Omega}}^{\top} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then, we have

$$\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^{\top}(\boldsymbol{V}_{\perp} \otimes \boldsymbol{U}_{\perp})(\boldsymbol{V}_{\perp} \otimes \boldsymbol{U}_{\perp})^{\top}\boldsymbol{S}_{\bar{\Omega}} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

and $\lambda_{\min}(\boldsymbol{H}) = 0$. *While Theorem 7.1 does not guarantee linear convergence of IHTSVD, one may realize that it is impossible to recover the last row of* $\boldsymbol{M}$ *in this case.*

Existing convergence analyses of algorithms for low-rank matrix completion often rely on standard assumptions, such as the incoherence of the underlying matrix $\boldsymbol{M}$ and the uniform randomness of the sampling pattern Ω [33]. Under these assumptions and a sample complexity bound on the number of observed entries s, linear convergence to a global solution can be guaranteed (see [106] for alternating minimization and [55] for IHT), with an upper bound on the rate of convergence $\rho < 1/2$. Our analysis, on the other hand, do not use the aforementioned assumptions but introduces a quantity that is fundamental to the problem in terms of optimization. By exploiting the local structure of the problem, we characterize the exact linear rate of local convergence of IHT. Similar to standard assumptions in prior works, the closed-form expression we obtained can be used to

determine sufficient conditions that ensure linear convergence. However, since our expression is exact, one can identify conditions that are potentially less stringent than existing conditions. For further details, we refer the interested readers to the Appendix 7.7.1.

7.3.2 Proof of Theorem 7.1

This section provides an overview of the proof of Theorem 7.1. We starts by formulating the recursion on the error matrix from the update (7.6) and the linearization of the rank-r projection:

Lemma 7.1. *Let us define the error matrix and its economy vectorized version, respectively, as*

$$\boldsymbol{E}^{(k)} = \boldsymbol{X}^{(k)} - \boldsymbol{M} \qquad \textit{and} \qquad \boldsymbol{e}^{(k)} = \boldsymbol{S}_{\bar{\Omega}}^{\top} \operatorname{vec}(\boldsymbol{E}^{(k)}).$$

Then, we have

$$\boldsymbol{E}^{(k+1)} = \mathcal{P}_{\bar{\Omega}}\big(\boldsymbol{E}^{(k)} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{E}^{(k)} \boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{E}^{(k)})\big), \tag{7.12}$$

$$\boldsymbol{e}^{(k+1)} = \big(\boldsymbol{I} - \boldsymbol{S}_{\bar{\Omega}}^{\top}(\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})\boldsymbol{S}_{\bar{\Omega}}\big)\boldsymbol{e}^{(k)} + \boldsymbol{r}\big(\boldsymbol{e}^{(k)}\big), \tag{7.13}$$

where $\boldsymbol{R}(\cdot)$ is the residual defined in Proposition 7.1 and

$$\boldsymbol{r}(\boldsymbol{e}) = \boldsymbol{S}_{\bar{\Omega}}^{\top} \operatorname{vec}\Big(\boldsymbol{R}\big(\overset{-1}{\operatorname{vec}}(\boldsymbol{S}_{\bar{\Omega}}\boldsymbol{e})\big)\Big) \quad \textit{for } \boldsymbol{e} \in \mathbb{R}^{n_1 n_2 - s}.$$

Here we recall that $\mathrm{vec}^{-1}(\cdot)$ *is the inverse vectorization operator such that* $(\mathrm{vec}^{-1} \circ \mathrm{vec})$ *is identity.*

Note that $\boldsymbol{E}^{(k)}$ belongs to the set of matrices supported in Ω and hence, $\left\|\boldsymbol{E}^{(k)}\right\|_F = \left\|\boldsymbol{e}^{(k)}\right\|_2$. Next, using the definition of the operator norm, one can obtain the following bound on the norm of the error matrix:

Lemma 7.2. *The Frobenius norm of the error matrix satisfies*

$$\left\|\boldsymbol{E}^{(k+1)}\right\|_F \leq \left(1 - \lambda_{\min}(\boldsymbol{H})\right) \left\|\boldsymbol{E}^{(k)}\right\|_F + \frac{c_1}{\sigma_r} \left\|\boldsymbol{E}^{(k)}\right\|_F^2 . \tag{7.14}$$

The nonlinear difference equation (7.14) has been well-studied in the stability theory of difference equation [15,166,216]. In fact, our theorem follows on applying Theorem 1 in [216] to (7.14), with $a_0 = \left\|\boldsymbol{E}^{(0)}\right\|_F$, $\alpha = 1 - \lambda_{\min}(\boldsymbol{H})$, and $\beta = c_1/\sigma_r$. The proofs of Lemmas 7.1 and 7.2 are given in Appendix 7.7.3.

7.3.3 IHT with Step Sizes Different than 1

Recall that IHTSVD is a special case of IHT with a unit step size. Thanks to the alternating-projection view in (7.4), the error $\boldsymbol{E}^{(k)} = \boldsymbol{X}^{(k)} - \boldsymbol{M}$ is guaranteed to be in the set of matrices supported in Ω, i.e., $\mathcal{P}_\Omega(\boldsymbol{E}^{(k)}) = \boldsymbol{E}^{(k)}$. Hence, the error analysis reduces from the space $\mathbb{R}^{n_1 \times n_2}$ for $\boldsymbol{E}^{(k)}$ to the space $\mathbb{R}^{n_1 n_2 - s}$ for $\boldsymbol{e}^{(k)} = \boldsymbol{S}_\Omega^\top \mathrm{vec}(\boldsymbol{E}^{(k)})$. While this appeal no longer holds for step sizes different than 1, one can one can follow a similar track to obtain an exact rate analysis in the general case.

Indeed, the linear convergence of IHT with a fixed step size $\eta > 0$ has recently studied in [217]. In particular, Vu *et. al.* proved that for $0 < \eta < 2/\|\boldsymbol{K}\|_2$, where $\boldsymbol{K} = \boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp \in \mathbb{R}^{r(n_1+n_2-r)\times r(n_1+n_2-r)}$ and $\boldsymbol{Q}_\perp \in \mathbb{R}^{n_1 n_2 \times r(n_1+n_2-r)}$ satisfies $\boldsymbol{Q}_\perp^\top \boldsymbol{Q}_\perp = \boldsymbol{I}_{r(n_1+n_2-r)}$ and $\boldsymbol{Q}_\perp \boldsymbol{Q}_\perp^\top = \boldsymbol{I}_{n_1 n_2} - \boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}$, the local linear convergence rate of IHT with a fixed step size η is given by

$$\rho_\eta = \max\{|1 - \eta\lambda_1(\boldsymbol{K})|, |1 - \eta\lambda_{r(m+n-r)}(\boldsymbol{K})|\}. \tag{7.15}$$

By comparing the two matrices $\boldsymbol{K} = \boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp \in \mathbb{R}^{r(n_1+n_2-r)\times r(n_1+n_2-r)}$ and $\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^\top (\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \boldsymbol{S}_{\bar{\Omega}} \in \mathbb{R}^{(n_1 n_2 - s)\times(n_1 n_2 - s)}$, we recognize that they share the same set of eigenvalues in the interval $[0, 1)$ while may only differ by the eigenvalues at 1. Thus, substituting $\eta = 1$ into (7.15) yields the same expression of the rate in (7.9).

It is also interesting to note that the optimal step size and the optimal convergence rate are given by [217]

$$\begin{aligned} \eta_{opt} &= \frac{2}{\lambda_1(\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp) + \lambda_{r(m+n-r)}(\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp)}, \\ \rho_{opt} &= 1 - \frac{2}{\kappa(\boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp) + 1}. \end{aligned} \tag{7.16}$$

While this results in faster convergence compared to IHTSVD, we focus on the latter in this work for simplicity and convenience of the analysis in the asymptotic matrix completion setting in Section 7.4. Further discussion on IHT with different step sizes is also given in Appendix 7.7.2.

7.4 Convergence of IHTSVD for Large-Scale Matrix Completion

In this section, we study the convergence of IHTSVD for large-scale matrix completion, a setting of practical interest in the rise of big data. Using recent results in random matrix theory, we show that, as its dimensions grow to infinity, the spectral distribution of $\boldsymbol{H}$ converges almost surely to a deterministic distribution with a bounded support. Consequently, we propose a large-scale asymptotic estimate of the linear convergence rate of IHTSVD that is a closed-form expression of the relative rank and the sampling rate.

7.4.1 Overview

We are interested in the asymptotic setting in which the size of $\boldsymbol{M}$ grows to infinity, i.e., $m = \min\{n_1, n_2\} \to \infty$. Let us assume that the ratio n_1/n_2 remains to be a non-zero constant as $m \to \infty$. In addition, we introduce two concepts that are the normalization of the degrees of freedom and the number of measurements:

Definition 7.5 (Relative rank)**.** *The rank r increases as $m \to \infty$ such that the relative rank remains to be a constant*

$$\rho_r = 1 - \sqrt{\big(1 - \frac{r}{n_1}\big)\big(1 - \frac{r}{n_2}\big)} \in (0, 1]. \tag{7.17}$$

Definition 7.6 (Sampling rate)**.** *The number of observations increases as $m \to \infty$*

such that the sampling rate remains to be a constant

$$\rho_s = \frac{s}{n_1 n_2} \in (0, 1]. \tag{7.18}$$

When $\rho_s < 1 - (1 - \rho_r)^2$, we recover the case in Remark 7.1 where the number of measurements is less than the degrees of freedom. As far as the local linear rate of IHTSVD is concerned, we only consider the case $\rho_s \geq 1 - (1 - \rho_r)^2$.

Remark 7.3. *When $r = m$, we have $\rho_r = 1$. Moreover, when $n_1 = n_2 = m$, the relative rank is exactly the ratio r/m. As can be seen below, the proposed definition of the relative rank incorporates both dimensions of $\boldsymbol{M}$ to enable the compact representation of ρ in terms of ρ_r and ρ_s.*

We are in position to state our result on the asymptotic behavior of the linear rate ρ in large-scale matrix completion:

Theorem 7.2 (Informal)**.** *For $\rho_s > 1 - (1 - \rho_r)^2$, the linear convergence rate ρ of IHTSVD approaches*

$$\rho_\infty = 1 - \left(\sqrt{(1 - \rho_r)^2 \rho_s} - \sqrt{\rho_r (2 - \rho_r)(1 - \rho_s)} \right)^2, \tag{7.19}$$

as $m \to \infty$.

Note that ρ_∞ is independent of the structure of the solution matrix $\boldsymbol{M}$ and the sampling set Ω. Moreover, it depends only on the relative rank and the sampling rate. Figure 7.1 depicts the contour plot of ρ_∞ as a function of ρ_r and ρ_s. It

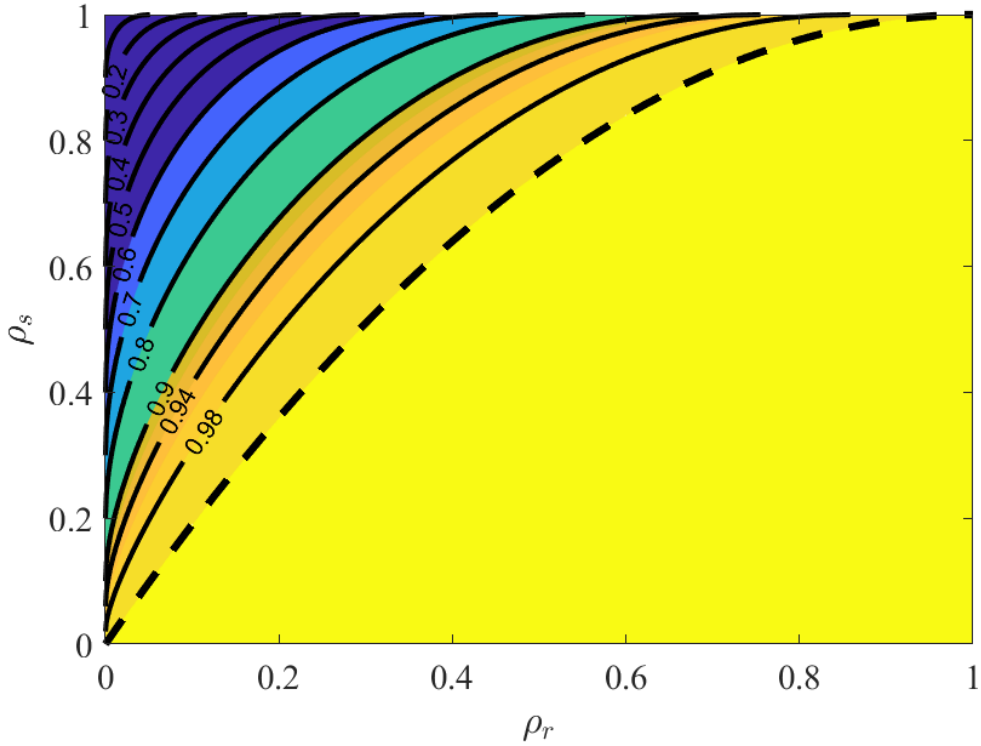


Figure 7.1: Contour plot of ρ_∞ as a 2-D function of ρ_r and ρ_s given by (7.19). The isoline at which $\rho_\infty = 1$ is represented by the dashed line. The white region below this isoline corresponds to the under-determined setting $\rho_s < 1 - (1 - \rho_r)$.

can be seen that for a fixed value of ρ_r, the asymptotic rate decreases towards 0 as the number of observed entries increases. This matches with the intuition that more information leads to faster convergence. Conversely, for a fixed value of ρ_s, the algorithm converges slower as the rank of the matrix increases, due to the increasing uncertainty (i.e., more degrees of freedom) in the set $\bar{\Omega}$. On the boundary where $\rho_s = 1 - (1 - \rho_r)^2$, there is no linear convergence predicted by our theory since $\rho_\infty = 1$. In this case, we recall that the number of observed entries equals the degrees of freedom of the problem.

Our technique relies on recent results in random matrix theory to exploit the special structure of $\boldsymbol{H}$. First, when n_1/n_2 remains constant, it holds that $n = n_1 n_2 \to \infty$ as $m \to \infty$. Then, $\boldsymbol{H}$ can be viewed as an element of a sequence of matrices

$$\boldsymbol{H}_n = \boldsymbol{W}_{pq}^n (\boldsymbol{W}_{pq}^n)^\top, \tag{7.20}$$

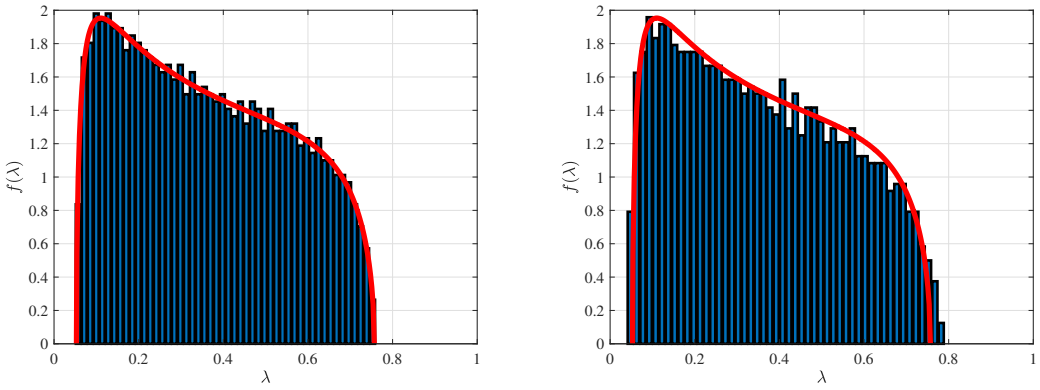
where $\boldsymbol{W}_{pq}^n \in \mathbb{R}^{pn_1n_2 \times qn_1n_2}$ is a truncation of the orthogonal matrix $\boldsymbol{W}^n = \boldsymbol{V}^{n_2} \otimes \boldsymbol{U}^{n_1}$, for $\boldsymbol{U}^{n_1}$ and $\boldsymbol{V}_\perp^{n_2}$ orthogonal matrices of dimensions $n_1 \times n_1$ and $n_2 \times n_2$, respectively, and

$$\begin{aligned} p &= \frac{n_1 n_2 - s}{n_1 n_2} = 1 - \rho_s, \\ q &= \frac{(n_1 - r)(n_2 - r)}{n_1 n_2} = (1 - \rho_r)^2. \end{aligned}$$

As n grows to infinity, we are interested in finding the limit (or even the limiting distribution) of the smallest eigenvalue of $\boldsymbol{H}_n$, which is a random truncation of the Kronecker product of two large dimensional semi-orthogonal matrices.

7.4.2 Truncations of Large Dimensional Orthogonal Matrices

Random matrix theory studies the asymptotic behavior of eigenvalues of matrices with entries drawn randomly from various matrix ensembles such as Gaussian orthogonal ensemble (GOE), Wishart ensemble, MANOVA ensemble [62]. The closest random matrix ensemble to our matrix ensemble $\{\boldsymbol{H}_n\}$ is the MANOVA



(a) (b)

Figure 7.2: Scaled histogram and the limiting ESD of $\boldsymbol{H}_n = \boldsymbol{W}_{pq}^n(\boldsymbol{W}_{pq}^n)^\top$, where $\boldsymbol{W}_{pq}^n$ is the $pn \times qn$ upper-left corner of an $n \times n$ orthogonal matrix $\boldsymbol{W}_n$, for $n = 10000$, $p = 0.16$, and $q = 0.36$. In (a), $\boldsymbol{W}_n$ is the orthogonal factor in the QR factorization of a 10000×10000 random matrix with $i.i.d$ standard normal entries. In (b), $\boldsymbol{W}_n = \boldsymbol{Q}_1 \otimes \boldsymbol{Q}_2$, where $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ are the orthogonal factors in the QR factorization of two independent 100×100 random matrices with $i.i.d$ standard normal entries. The histograms with 50 bins (blue) are scaled by a factor of $1/pnw$, where w is the bin width. The limiting ESD (red) is generated by (7.22). It can be seen that the histogram in (a) match the limiting ESD better than the histogram in (b).

ensemble in which truncations of large dimensional Haar orthogonal matrices are considered. Here we recall that the Haar measure provides a uniform distribution over the set of all $n \times n$ orthogonal matrices $\mathbb{O}(n)$. Indeed, it is a unique translation-invariant probability measure on $\mathbb{O}(n)$. If we assume that the matrix $\boldsymbol{M}$ follows a random orthogonal model [33], then $\boldsymbol{U}_\perp$ and $\boldsymbol{V}_\perp$ are essentially sub-matrices of Haar orthogonal matrices in $\mathbb{O}(n_1)$ and $\mathbb{O}(n_2)$, respectively, and $\{\boldsymbol{H}_n\}$ is a truncation of the Kronecker product of two Haar orthogonal matrices.

There has been certain theoretical work on truncations of Haar invariant matrices in the literature. In 1980, Wachter [219] established the limiting distribution of the eigenvalues in the MANOVA ensemble. Later on, the density function of the eigenvalues of such matrix has been shown to be the same as that of a Jacobi matrix [37, 47, 68]. Shortly afterward, Johnstone proved the Tracy-Widom behavior of the largest eigenvalue in [111]. More recently, Farrell and Nadakuditi relaxed the constraint on the uniform (Haar) distribution of the orthogonal matrix considered the Kronecker products of Haar-distributed orthogonal matrices, which is similar to our matrix completion setting in this chapter. The authors showed that the limiting density of their truncations remains the same as the original case without Kronecker products. Further results on the eigenvalues distribution of truncations of Haar orthogonal matrices were also given in [57, 108, 239]. To the best of our knowledge, no result has been shown for the limiting behavior of the smallest eigenvalue of random MANOVA matrices.

In our context, we leverage the recent result in [171], which assumes the randomness on the truncation rather than the orthogonal matrix. This variant, while

differs from the classic MANOVA ensemble in random matrix theory, is well-suited to the setting of matrix completion. Let us begin with the following definition of the empirical spectral distribution:

Definition 7.7. *Let $\boldsymbol{H}_n$ be an $n \times n$ real symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. The **empirical spectral distribution (ESD)** of $\boldsymbol{H}_n$, denoted by $\mu_{\boldsymbol{H}_n}$, is the probability measure which puts equal mass at each of the eigenvalues of $\boldsymbol{H}_n$:*

$$\mu_{\boldsymbol{H}_n} \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i},$$

where δ_λ is the Dirac mass at λ.

Next, we define the concepts of a sequence of row sub-sampled matrices and the concentration property:

Definition 7.8. *For each $n \in \mathbb{N}^+$, consider the $n \times qn$ matrix $\boldsymbol{W}_q^n = [\boldsymbol{w}_1^n, \ldots, \boldsymbol{w}_n^n]^\top$, where $\boldsymbol{w}_i^n \in \mathbb{R}^{qn}$ and q is a constant in $(0, 1)$. Let P_n be a pn-permutation of $[n]$ selected uniformly at random, for p is a constant in $(0, 1)$, and $\boldsymbol{W}_{pq}^n \in \mathbb{R}^{pn \times qn}$ be the random matrix obtained by selecting the corresponding set of pn rows from $\boldsymbol{W}_q^n$. Then, the sequence $\{\boldsymbol{W}_q^n\}_{n \in \mathbb{N}^+}$ is called **a sequence of q-tall matrices**, and the sequence $\{\boldsymbol{W}_{pq}^n\}_{n \in \mathbb{N}^+}$ is called **a sequence of row sub-sampled matrices** of $\{\boldsymbol{W}_q^n\}_{n \in \mathbb{N}^+}$.*

Definition 7.9. *Given the setting in Definition 7.8, for each $j \in P_n$, denote*

$P_n^j = P_n \setminus \{j\}$. *In addition, for* $z \in \mathbb{C}$, *define*

$$\boldsymbol{R}_j(z) = \Big(\sum_{i \in P_n^j} \boldsymbol{w}_i^n (\boldsymbol{w}_i^n)^\top - z \boldsymbol{I}_{qn} \Big)^{-1}.$$

Then, the sequence $\{\boldsymbol{W}_q^n\}_{n \in \mathbb{N}^+}$ *is* ***concentrated*** *if and only if for any* $j \in P_n$ *and* $z \in \mathbb{C}$, *we have*

$$(\boldsymbol{w}_j^n)^\top \boldsymbol{R}_j(z) \boldsymbol{w}_j^n - \mathbb{E}_{j|P_n^j} \big[(\boldsymbol{w}_j^n)^\top \boldsymbol{R}_j(z) \boldsymbol{w}_j^n \big] \xrightarrow{p} 0. \tag{7.21}$$

In the following, we consider examples of sequences of matrices that are concentrated, as well as an example of the sequence of incoherent matrices that are **not** concentrated.

Example 7.1. *Random settings:*

1. *The sequence of* q*-tall matrices* $\{\boldsymbol{A}_q^n\}_{n \in \mathbb{N}^+}$, *where the entries of* $\boldsymbol{A}_q^n$ *are i.i.d* $\mathcal{N}(0, 1/n)$, *is concentrated.*

2. *The sequence* $\{\boldsymbol{B}_q^n \otimes \boldsymbol{C}_q^n\}_{n \in \mathbb{N}^+}$, *where* $\{\boldsymbol{B}_q^n\}_{n \in \mathbb{N}^+}$ *and* $\{\boldsymbol{C}_q^n\}_{n \in \mathbb{N}^+}$ *are two sequences of* q*-tall matrices whose entries are i.i.d* $\mathcal{N}(0, 1/n)$, *is also concentrated.*

We provide the detailed explanation of this example in Appendix 7.7.4.

Example 7.2. *Deterministic settings:*

1. *The sequence of* q*-tall matrices* $\{\boldsymbol{D}_q^n\}_{n \in \mathbb{N}^+}$, *where the entries of* $\boldsymbol{D}_q^n$ *are all* 1, *is concentrated.*

2. *The sequence of* $1/2$*-tall matrices* $\{\boldsymbol{E}_q^n\}_{n\in\mathbb{N}^+}$ *where*

$$\boldsymbol{E}_q^n = \begin{bmatrix} 0.6\sqrt{\frac{2}{n}}\mathtt{H}_{n/2} \\ 0.8\sqrt{\frac{2}{n}}\mathtt{H}_{n/2} \end{bmatrix},$$

for $\mathtt{H}_{n/2}$ *being a Hadamard matrix of order* $n/2$ *[95], is not concentrated. On the other hand, one can verify that* $\boldsymbol{E}_q^n$ *is* μ*-incoherent, for*

$$\mu = \left\|0.8\sqrt{2/n}\mathtt{H}_{n/2}\right\|_F^2 \frac{n}{n/2} = 1.28.$$

Thus, the concentration assumption considered in this chapter is stronger than the widely-used incoherence assumption.

With these definitions in place, we now state the result on the limiting ESD of a truncation of orthogonal matrices. To fit our matrix completion setting in this chapter, we rephrase the result in [171] to the case of row sub-sampled semi-orthogonal matrices (as opposed to column sub-sampled semi-orthogonal matrices in the aforementioned paper).

Proposition 7.2 (Rephrased from [171])**.** *Let* $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$ *be a sequence of* q*-tall matrices that is concentrated. In addition, assume that* $\boldsymbol{W}_q^n$ *is semi-orthogonal for all* $n \in N^+$*, i.e.,* $(\boldsymbol{W}_q^n)^\top \boldsymbol{W}_q^n = \boldsymbol{I}_{qn}$*. Let* $\{\boldsymbol{W}_{pq}^n\}_{n\in\mathbb{N}^+}$ *be a sequence of row sub-sampled matrices of* $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$*. Then, as* $n \to \infty$*, the ESD of* $\boldsymbol{H}_n = \boldsymbol{W}_{pq}^n(\boldsymbol{W}_{pq}^n)^\top$

converges almost surely to the deterministic distribution μ_{pq} *such that*

$$d\mu_{pq} = \Big(1 - \frac{q}{p}\Big)_+ \delta(x)dx + \Big(\frac{p+q-1}{p}\Big)_+ \delta(x-1)dx + \frac{\sqrt{(\lambda^+ - x)(x - \lambda^-)}}{2\pi p x(1-x)} \mathbb{I}[\lambda^- \le x \le \lambda^+]dx, \tag{7.22}$$

where δ *is the Dirac delta function and*

$$\lambda^\pm = \big(\sqrt{q(1-p)} \pm \sqrt{p(1-q)}\big)^2.$$

The proposition asserts that the limiting ESD of $\boldsymbol{H}_n$ exists and depends only on the row ratio p and the column ratio q, provided that $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$ is concentrated. We note that the distribution μ_{pq} is exactly the same as the limiting distribution of the MANOVA ensemble. Indeed one can show that the MANOVA ensemble as a special case of the concentrated sequence:

Lemma 7.3. *Let* $\boldsymbol{W}^n$ *be a Haar-distributed orthogonal matrix in* $\mathbb{O}(n)$ *and* $\boldsymbol{W}_q^n$ *be the semi-orthogonal matrices obtained from any* qn *(for* $q \in (0,1)$*) columns of* $\boldsymbol{W}^n$*. Then the sequence* $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$ *is concentrated.*

Furthermore, the Kronecker product of two Haar-distributed orthogonal matrices also possesses the concentration property:

Lemma 7.4. *Let* $\boldsymbol{U}^{n_1}$ *and* $\boldsymbol{V}^{n_2}$ *be Haar-distributed orthogonal matrices in* $\mathbb{O}(n_1)$ *and* $\mathbb{O}(n_2)$*, respectively. Define* $\boldsymbol{U}_{q_1}^{n_1}$ *and* $\boldsymbol{V}_{q_2}^{n_2}$ *as the semi-orthogonal matrices obtained from any* q_1 *and* q_2 *(for* $q_1, q_2 \in (0,1)$*) columns of* $\boldsymbol{U}^{n_1}$ *and* $\boldsymbol{V}^{n_2}$*, respectively. Then the sequence* $\{\boldsymbol{W}_q^n = \boldsymbol{U}_{q_1}^{n_1} \otimes \boldsymbol{V}_{q_2}^{n_2}\}_{n\in\mathbb{N}^+}$ *(with* $q = q_1 q_2$*) is concentrated.*

Lemmas 7.3 and 7.4 are immediate consequences of Lemma 3.1 in [63], so we omit the proof of these lemmas here.

7.4.3 Proposed Estimation of ρ

In order to apply Proposition 7.2 to our matrix completion setting, we recall that $\boldsymbol{W}_{pq}^n$ can be viewed as the n-th element of a sequence of row sub-sampled matrices of $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$, where $\boldsymbol{W}_q^n = \boldsymbol{V}_\perp^{n_2} \otimes \boldsymbol{U}_\perp^{n_1}$. If the sequence $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$ is concentrated, then (7.22) holds for $p = 1-\rho_s$ and $q = (1-\rho_r)^2$. Therefore, one might expect that the smallest eigenvalue of $\boldsymbol{H}_n = \boldsymbol{W}_{pq}^n (\boldsymbol{W}_{pq}^n)^\top$ converges to

$$\lambda^- = \left(\sqrt{q(1-p)} - \sqrt{p(1-q)}\right)^2.$$

Thus, by Theorem 7.1, the convergence rate ρ converges to $1-\lambda^-$. The following theorem is an immediate application of Proposition 7.2 to our large-scale matrix completion setting:

Theorem 7.3. *As $m \to \infty$, assume that $\boldsymbol{M}$ is generated in a way that the Kronecker product $\boldsymbol{W}_q^n = \boldsymbol{V}_\perp^{n_2} \otimes \boldsymbol{U}_\perp^{n_1}$ forms a sequence of semi-orthogonal matrices that is concentrated. Then, provided $\rho_s \geq 1-(1-\rho_r)^2$, the ESD $\mu_{\boldsymbol{H}_n}$ converges almost surely to the deterministic distribution $\mu_{\rho_r\rho_s}$ such that*

$$d\mu_{\rho_r\rho_s} = \left(\frac{(1-\rho_r)^2-\rho_s}{1-\rho_s}\right)_+ \delta(x-1)dx + \frac{\sqrt{(\lambda^+-x)(x-\lambda^-)}}{2\pi(1-\rho_s)x(1-x)}\mathbb{I}[\lambda^- \leq x \leq \lambda^+]dx, \tag{7.23}$$

where $\lambda^{\pm} = \left(\sqrt{(1-\rho_r)^2\rho_s} \pm \sqrt{\rho_r(2-\rho_r)(1-\rho_s)}\right)^2$.

While the theorem claims the convergence of the spectral distribution of $\boldsymbol{H}$, it does not imply the convergence of its smallest eigenvalue to λ^-. In fact, it is not trivial to prove this fact. We leave the following as an open question for future work.

Conjecture 7.1. *Assume the same setting as in Theorem 7.3. As $m \to \infty$, the linear rate ρ defined in (7.9) converges almost surely to p_∞ defined in (7.19).*

7.5 Numerical Results

In this section, we provide numerical results to verify the exact linear convergence rate of IHTSVD in (7.9) and to compare this analytical rate with the asymptotic rate in (7.19) in large-scale settings.

7.5.1 Analytical Rate versus Empirical Rate

In this experiment, we verify the analytical expression of linear convergence rate of IHTSVD by comparing it with the empirical rate obtained by measuring the decrease in the norm of the error matrix. Our goal is to demonstrate that they agree in various settings of ρ_r and ρ_s.

Data generation. We first set the dimensions $n_1 = 50$ and $n_2 = 40$. Next, for each r in $\{1, 2, \ldots, 12\}$, we generate the rank-r matrix $\boldsymbol{M}$ as follows. We construct the random orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ by (i) generating a $n_1 \times n_2$ random matrix

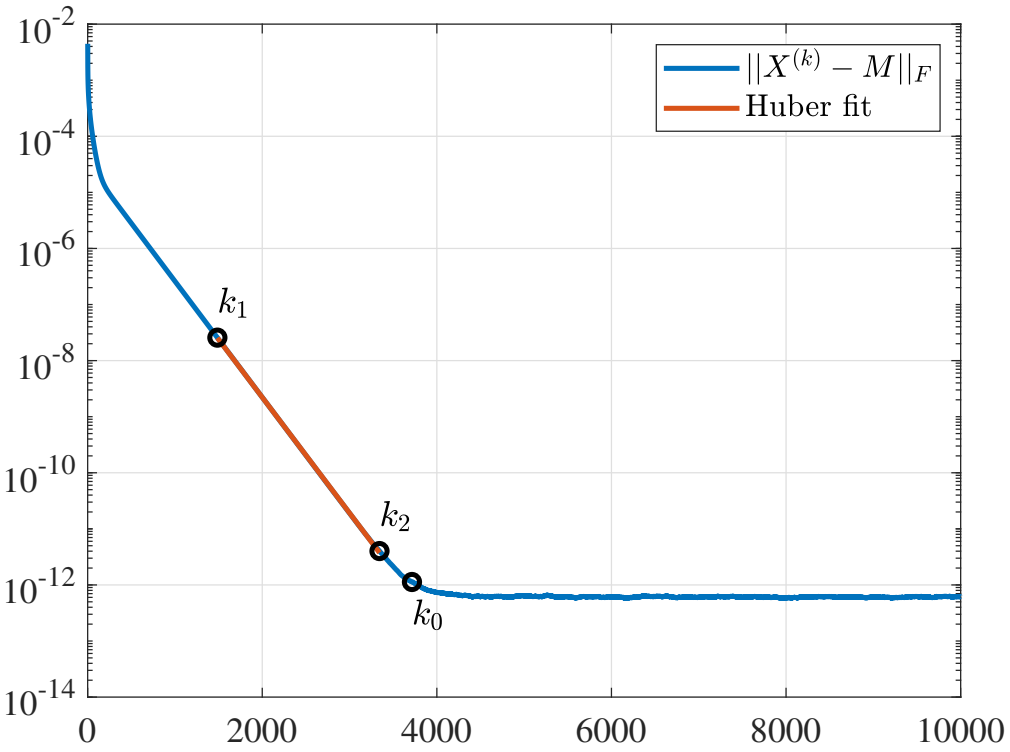
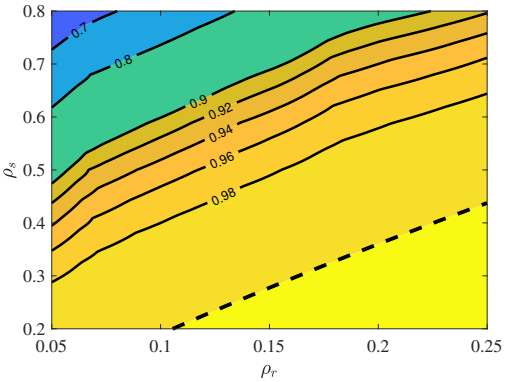
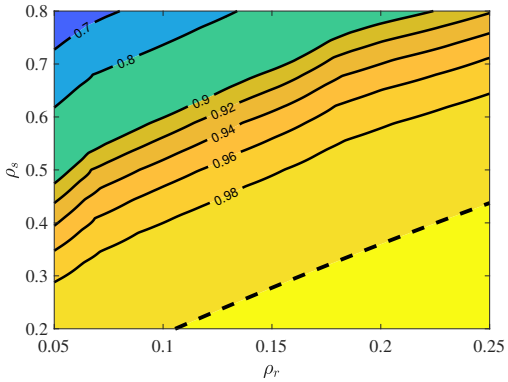


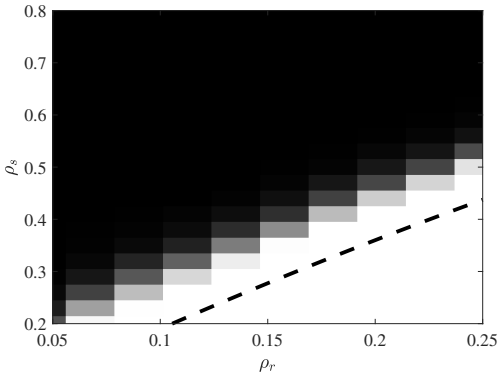
Figure 7.3: Estimation of the empirical rate using the error sequence $\{\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\|_F\}_{k=k_1}^{k_2}$. Due to the numerical error below 10^{-12}, we need to identify the 'turning point' at k_0 and then set $k_1 = \lfloor 0.4k_0 \rfloor$ and $k_2 = \lfloor 0.9k_0 \rfloor$.



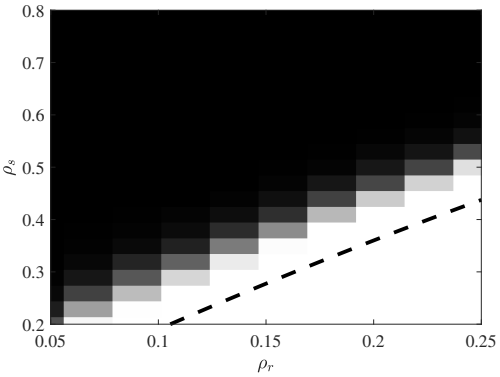
(a) Analytical rate



(b) Empirical rate



(c) Probability of linear convergence (analytical)



(d) Probability of linear convergence (empirical)

Figure 7.4: Comparison between the analytical rate and the empirical rate of convergence of IHTSVD in various matrix completion settings for $n_1 = 50, n_2 = 40$. (a) Contour plot of the analytical rate as a function of ρ_r and ρ_s. (b) Contour plot of the empirical rate as a function of ρ_r and ρ_s. (c) Probability of linear convergence based on the analytical rate. (d) Probability of linear convergence based on the empirical rate. The black color corresponds to linear convergence, while the white color corresponds to no linear convergence. In each plot, the data is interpolated based on a 12×21 grid over ρ_r and ρ_s, in which the value of each point is evaluated by 1000 runs. Additionally, a dashed line is included to indicate the line $1 - \rho_s = (1 - \rho_r)^2$. It can be seen that there is a perfect match between the analytical rate and the empirical rate.

whose entries are $i.i.d$ normally distributed $\mathcal{N}(0,1)$ and (ii) performing the singular value decomposition of the resulting matrix. The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are comprised of the corresponding left and right singular vectors. Then, the rank-r matrix $\boldsymbol{M}$ is generated by taking the product $\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top$, where $\boldsymbol{\Sigma}_1 = \text{diag}(r, r-1, \ldots, 1)$ and $\boldsymbol{U}_1, \boldsymbol{V}_1$ are the first r columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Finally, for each s in the linearly spaced set $\{0.2n, 0.23n, 0.26n, \ldots, 0.8n\}$, we create the 1000 different sampling sets, each of them is obtained by generating a random permutation of the set $[n]$ and then selecting the first s elements of the permutation. Thus, we obtain a 12×21 grid based on the values of r and s such that (i) grid points corresponding to the same rank r share the same underlying matrix $\boldsymbol{M}$; (ii) each point on the grid corresponds to 1000 different sampling sets.

Estimating Analytical Rate and Empirical Rate. We calculate the analytical rate for each aforementioned setting of $\boldsymbol{M}$ and Ω using (7.9). Due to numerical errors in computing small eigenvalues, we need to set all the resulting rates that are greater than 1 to 1, indicating there is no linear convergence in such cases. For the calculation of the empirical rate, we run Algorithm 7.1 in the same setting with $K = 10000$ iterations. The initial point $\boldsymbol{X}^{(0)}$ is obtained by adding $i.i.d.$ normally distributed noise with standard deviation $\sigma = 10^{-4}$ to the entries of $\boldsymbol{M}$. Here we note that σ is chosen to be small for two reasons: (i) for large matrices, even small σ for individual entry can add up to a large error on the entire matrix; and (ii) while the cost of computing $\lambda_{\min}$ (and hence, the region of convergence) is prohibitively expensive for large matrices, choosing small σ empirically guarantees the initialization is inside the region of convergence. Then, we record the error sequence

$\{\left\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\right\|_F\}_{k=1}^{K}$ and determine if the algorithm converges linearly to $\boldsymbol{M}$ by checking whether there exists $\hat{K} \leq K$ such that $\left\|\boldsymbol{X}^{(\hat{K})} - \boldsymbol{M}\right\|_F < \epsilon \left\|\boldsymbol{X}^{(0)} - \boldsymbol{M}\right\|_F$, for $\epsilon = 10^{-8}$. If the relative error is above ϵ, we set the empirical rate to 1 to indicate that the algorithm does not converge linearly. However, it is important to note that this heuristic does not perfectly detect linear convergence since it overlooks the case in which the linear rate is extremely close to 1 and it requires more than $K = 10000$ iterations to reach a relative error below ϵ. As can be seen later, to compromise this computational limit, we resort to setting the analytical rate that is greater than 0.998 to 1 when making a comparison between the analytical rate and the empirical rate[4]. In case the relative error is less than ϵ, we terminate the algorithm at the $\hat{K}$-th iteration (early stop) and perform an estimation on the error sequence $\{\left\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\right\|_F\}_{k=1}^{\hat{K}}$ to obtain the empirical rate.

After obtaining the analytical rate and the empirical rate over the 2-D grid, we report the result in the contour plots of the rate as a function of ρ_r and ρ_s in Fig. 7.4-(a) and (b). Since our original grid is non-uniform, we perform a scattered data interpolation, which uses a Delaunay triangulation of the scattered sample points to perform interpolation [7], to evaluate the rate over a 1001×1001 uniform grid based on ρ_r and ρ_s. Due to the aforementioned limitation of estimating the empirical rate, we apply a threshold of 0.998 to both of the interpolated data for the analytical rate and the empirical rate, setting any value above the threshold to 1. In addition, we calculate the probability of linear convergence for the analytical

[4]Substituting $\epsilon = 10^{-8}$ and $N(\epsilon) = 10000$ into (7.10) and assuming the constant c_2 is negligible, we obtain $\lambda_{\min}(\boldsymbol{H}) \approx 1.8 \times 10^{-3}$.

rate and the empirical rate and visualize the result in Fig. 7.4-(c) and (d). As mentioned, we use a threshold of 0.998 to determine the linear convergence.

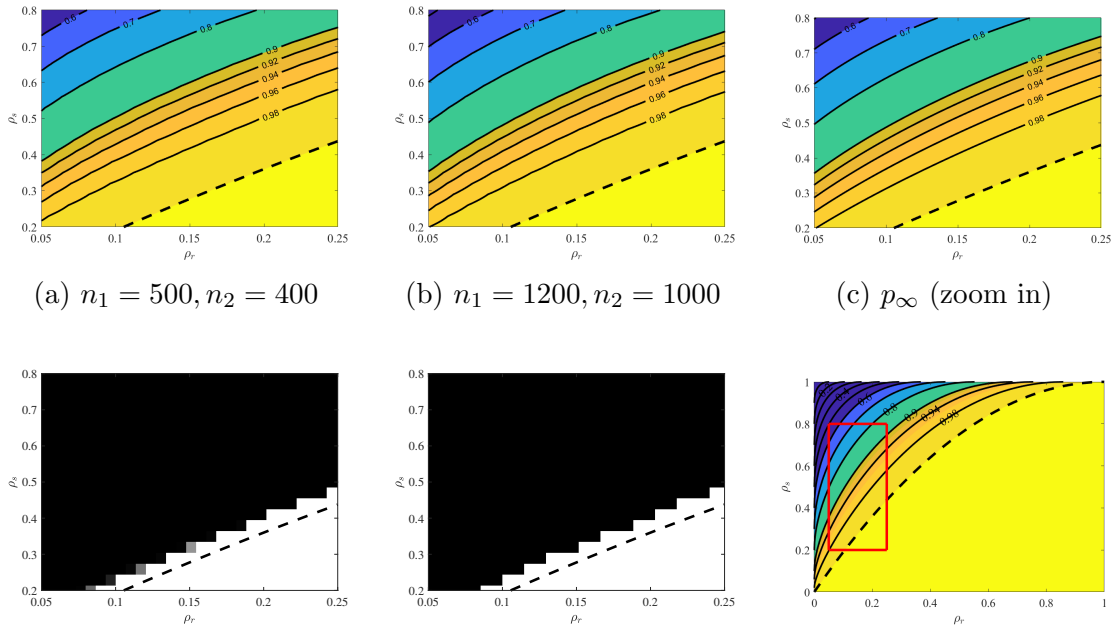
Results. Given the values of the analytical rate and the empirical rate of 1000 matrix completion settings for each point on the 12×21 grid, the mean squared difference between the two rates in our experiment is 2.9659×10^{-5}. Figures 7.4 illustrate the similarity between the analytical rate and the empirical rate evaluated under various settings of matrix completion. In both Fig. 7.4-(a) and Fig. 7.4-(b), we observed a matching behavior as in Fig. 7.1: smaller rank and more observation result in faster linear convergence of IHTSVD. However, the contour lines in Fig. 7.4 are not as smooth as those with asymptotic behavior in Fig. 7.1 due to the large variance when n_1 and n_2 are relatively small. On the other hand, it can be seen in Fig. 7.4-(c) and Fig. 7.4-(d) that there is a linear-convergence area (black) above the boundary line at $1 - \rho_s = (1 - \rho_r)^2$ and a no-linear-convergence area (white) below the boundary line. The transition area (gray) near above the boundary line corresponds to the settings in which some sampling sets result in $\lambda_{\min}(\boldsymbol{H}) = 0$ (no linear convergence) while some other sampling sets result in non-zero $\lambda_{\min}(\boldsymbol{H})$ (linear convergence). Note that in order to obtain the analytical rate, we need to compute the smallest eigenvalue of a $(n - s) \times (n - s)$ matrix, which is computationally expensive for large $n = n_1 n_2$. In particular, when $s = \mathcal{O}(n)$, the cost of computing the analytical rate is $\mathcal{O}(n^2)$. On the other hand, the empirical rate offers an alternative but more efficient way to estimate the convergence rate via running Algorithm 7.1 whose computational complexity per iteration is $O(nr)$.

7.5.2 Non-asymptotic Rate versus Asymptotic Rate

In this experiment, we compare the asymptotic rate given in Theorem 7.3 with the convergence rate of IHTSVD for large-scale matrix completion. For convenience, we refer the later as the non-asymptotic rate. As mentioned in the previous subsection, we use the empirical rate instead of the analytical rate to estimate the non-asymptotic rate due to the computational efficiency.

Data generation. We consider two settings of (n_1, n_2): $n_1 = 500, n_2 = 400$ and $n_1 = 1200, n_2 = 1000$. Similar to the previous experiment, we generate $\boldsymbol{M}$ and Ω based on a 2-D grid over r and s. While the values of s are still selected from the set $\{0.2n, 0.23n, 0.26n, \ldots, 0.8n\}$, the values of r are chosen differently for each setting of (n_1, n_2). In particular, for $n_1 = 500, n_2 = 400$, we select the values of r from the linearly spaced set $\{1, 4, 7, \ldots, 118\}$. For $n_1 = 1200, n_2 = 1000$, we select the values of r from the linearly spaced set $\{1, 9, 17, \ldots, 297\}$. Therefore, in the former setting, the grid size is 40×21, while in the later setting, the grid size is 38×21.

Implementation. The calculation of the empirical rate is the same as the previous experiment. For computational efficiency, we omit the points on the grid that are below the boundary line, i.e., $s < (n_1 + n_2 - r)r$, since it is evident that there is no linear convergence at such points. No analytical rate is given in this experiment because calculating the smallest eigenvalue of a $(n-s)\times(n-s)$ matrix is computationally expensive for large n_1 and n_2. On the other hand, the contour plot of the asymptotic rate is easy to obtained using (7.19).



(a) $n_1 = 500, n_2 = 400$ (b) $n_1 = 1200, n_2 = 1000$ (c) p_∞ (zoom in)

(d) $n_1 = 500, n_2 = 400$ (e) $n_1 = 1200, n_2 = 1000$ (f) p_∞ (zoom out)

Figure 7.5: Comparison between the empirical rate and the asymptotic rate of convergence of IHTSVD in various matrix completion settings. (a) Contour plot of the empirical rate as a function of ρ_r and ρ_s for $n_1 = 500, n_2 = 400$. (b) Contour plot of the empirical rate as a function of ρ_r and ρ_s for $n_1 = 1200, n_2 = 1000$. (c) Contour plot of the asymptotic rate as a function of ρ_r in range $[0.05, 0.25]$ and ρ_s in range $[0.2, 0.8]$. (d) Probability of linear convergence based on the empirical rate in (a). (e) Probability of linear convergence based on the empirical rate in (b). (The black color corresponds to linear convergence, while the white color corresponds to no linear convergence) (f) Contour plot of the asymptotic rate as a function of ρ_r in range $[0, 1]$ and ρ_s in range $[0, 1]$. The red solid rectangular corresponds to the zoom-in region in (c). In each plot, the data is interpolated based on a 2-D grid over ρ_r and ρ_s, in which the value of each point is evaluated by 100 runs. Additionally, a dashed line is included to indicate the line $1 - \rho_s = (1 - \rho_r)^2$.

Results. Fig. 7.5 compares the non-asymptotic rate and the asymptotic rate in various settings of ρ_r and ρ_s. As n_1 and n_2 increase, we observe that the contour lines of the non-asymptotic rate become smoother and approach those of the asymptotic rate. Compared with Fig. 7.4, it can also be seen that the isoline of the same value shifts down towards the boundary line as n_1 and n_2 increases.

7.6 Conclusion and Future Work

In this chapter, we established a closed-form expression of the linear convergence rate of an iterative hard thresholding method for solving matrix completion. We also identified the local region around the solution that guarantees the convergence of the algorithm. Furthermore, in large-scale setting, we leveraged the result from random matrix theory to offer a simple estimation of the asymptotic convergence rate in practice. Under certain assumption, we showed that the convergence rate of IHTSVD converges almost surely to our proposed estimate.

In future work, we would like to extend our local convergence analysis to other IHT methods with different step size, e.g., SVP [104] and accelerated IHT [213,214]. Moreover, it would be interesting to study the non-asymptotic behavior of the convergence rate in large-scale settings. Finally, we believe the technique presented in this chapter can be applied to study the local convergence of other non-convex methods such as alternating minimization [106] and gradient descent [199].

7.7 Appendix

7.7.1 Comparison to prior results

In our main theorem, the rate of convergence depends on

$$\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^{\top}(\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})\boldsymbol{S}_{\bar{\Omega}} \in \mathbb{R}^{(n_1 n_2 - s)\times(n_1 n_2 - s)},$$

where $\boldsymbol{S}_{\bar{\Omega}} \in \mathbb{R}^{n_1 n_2 \times (n_1 n_2 - s)}$ is the selection matrix corresponding to the complement set $\bar{\Omega}$. $P_{\boldsymbol{U}_\perp}$ and $P_{\boldsymbol{V}_\perp}$ are the projections onto the left and right null spaces of $\boldsymbol{M}$. Viewing $\boldsymbol{H}$ as a function of $\boldsymbol{M}$ and Ω, let us consider the set

$$\mathcal{S} = \{(\boldsymbol{X}, \Omega) \mid \boldsymbol{H}(\boldsymbol{X}, \Omega) \text{ is full rank } \}.$$

In the following, we show that our proposed set $\mathcal{S}$ contains the set of incoherent matrices and uniform sampling patterns. In other words, if $\boldsymbol{M}$ is incoherent and Ω is a uniform sampling, then $(\boldsymbol{M}, \Omega) \in \mathcal{S}$ *w.h.p.* First, we highlight the fact that the invertibility of $\boldsymbol{H}$ is related to the injectivity of the sampling operator restricted to $T_{\boldsymbol{M}}(\mathcal{M}_{\leq r})$ - the tangent space T to $\mathcal{M}_{\leq r} = \{\boldsymbol{X} \in R^{n_1 \times n_2} \mid \text{rank}(\boldsymbol{X}) \leq r\}$ at $\boldsymbol{M}$. In particular, recall that this operator is of the form $\mathcal{A}_{\Omega T} = \mathcal{P}_\Omega \mathcal{P}_T$, where $\mathcal{P}_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is the orthogonal projector onto the indices in Ω and $\mathcal{P}_T : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is the orthogonal projection onto T

(see [33]-Eqn. 3.5)

$$\mathcal{P}_T(\boldsymbol{X}) = \boldsymbol{P_U X} + \boldsymbol{X P_V} - \boldsymbol{P_U X P_V} = \boldsymbol{X} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{X} \boldsymbol{P}_{\boldsymbol{V}_\perp},$$

for all $\boldsymbol{X} \in \mathbb{R}^{m \times n}$. Using vectorization, one can show that

$$\text{vec}(\mathcal{P}_\Omega(\boldsymbol{X})) = \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \text{vec}(\boldsymbol{X}) = (\boldsymbol{I}_{n_1 n_2} - \boldsymbol{S}_{\bar{\Omega}} \boldsymbol{S}_{\bar{\Omega}}^\top) \text{vec}(\boldsymbol{X})$$

and

$$\begin{aligned} \text{vec}(\mathcal{P}_T(\boldsymbol{X})) &= (\boldsymbol{I}_{n_1 n_2} - \boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \text{vec}(\boldsymbol{X}) \\ &= \boldsymbol{Q}_\perp \boldsymbol{Q}_\perp^\top \text{vec}(\boldsymbol{X}), \end{aligned}$$

where $\boldsymbol{Q}_\perp \in \mathbb{R}^{n_1 n_2 \times r(n_1+n_2-r)}$ is the basis of $T_{\boldsymbol{M}}(\mathcal{M}_{\leq r})$, i.e., $\boldsymbol{Q}_\perp \boldsymbol{Q}_\perp^\top = \boldsymbol{I}_{n_1 n_2} - \boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}$. Therefore, the eigenvalues of the operator $\mathcal{A}_{\Omega T}^* \mathcal{A}_{\Omega T} = \boldsymbol{P}_T \boldsymbol{P}_\Omega \boldsymbol{P}_T : \mathbb{R}^{n_1 n_2} \to \mathbb{R}^{n_1 n_2}$ restricted to $T_{\boldsymbol{M}}(\mathcal{M}_{\leq r})$ are the same as those of the $r(n_1+n_2-r) \times r(n_1+n_2-r)$ matrix

$$\hat{\boldsymbol{H}} = \boldsymbol{Q}_\perp^\top (\boldsymbol{I}_{n_1 n_2} - \boldsymbol{S}_{\bar{\Omega}} \boldsymbol{S}_{\bar{\Omega}}^\top) \boldsymbol{Q}_\perp = \boldsymbol{I}_{r(n_1+n_2-r)} - \boldsymbol{Q}_\perp^\top \boldsymbol{S}_{\bar{\Omega}} \boldsymbol{S}_{\bar{\Omega}}^\top \boldsymbol{Q}_\perp.$$

Now representing $\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^\top (\boldsymbol{I}_{r(n_1+n_2-r)} - \boldsymbol{Q}_\perp \boldsymbol{Q}_\perp^\top) \boldsymbol{S}_{\bar{\Omega}} = \boldsymbol{I}_{n_1 n_2 - s} - \boldsymbol{S}_{\bar{\Omega}}^\top \boldsymbol{Q}_\perp \boldsymbol{Q}_\perp^\top \boldsymbol{S}_{\bar{\Omega}}$, it can be showed that $\boldsymbol{H}$ and $\hat{\boldsymbol{H}}$ share the same set of eigenvalues except those at 1. Equivalently, the injectivity of $\mathcal{A}_{\Omega T}$ restricted to $T_{\boldsymbol{M}}(\mathcal{M}_{\leq r})$ implies the invertibility of $\boldsymbol{H}$. Second, we recall the so-called result from Candes and Recht that the

operator $\mathcal{A}_{\Omega T}$ is *most likely* injective when restricted to $T_{\boldsymbol{M}}(\mathcal{M}_{\leq r})$. Specifically, Eqn. (4.11) in [33] states that if Ω is sampled according to the Bernoulli model with probability $p \approx s/n_1 n_2$ and the solution $\boldsymbol{M}$ is a rank-r matrix satisfying μ-coherent property, then for all $\boldsymbol{Y} \in \mathbb{R}^{n_1 \times n_2}$:

$$\begin{aligned}(1-\tau)p \left\|\boldsymbol{P}_T(\boldsymbol{Y})\right\|_F \leq \left\|(\boldsymbol{P}_T \boldsymbol{P}_\Omega \boldsymbol{P}_T)(\boldsymbol{Y})\right\|_F \\ \leq (1+\tau)p \left\|\boldsymbol{P}_T(\boldsymbol{Y})\right\|_F, \qquad w.h.p.,\end{aligned}$$

where τ is an arbitrarily small constant such that $C_R\sqrt{\frac{\mu n r \log n}{s}} \leq \tau < 1$, [5] for $n = \max(n_1, n_2)$. Finally, translating this into our context, we can show that under the same assumptions (uniform sampling and incoherence property) and *w.h.p.*, the matrix $\boldsymbol{H}$ is full rank with the property

$$\|\boldsymbol{H}\boldsymbol{x}\| \geq (1-\tau)p \|\boldsymbol{x}\|, \quad \forall \boldsymbol{x} \in \mathbb{R}^{n_1 n_2 - s}.$$

This implies $\lambda_{\min}(\boldsymbol{H}) \geq (1-\tau)p > 0$. We conclude that if $\boldsymbol{M}$ is incoherence and Ω is a uniform sampling, then $(\boldsymbol{X}, \Omega) \in \mathcal{S}$ *w.h.p.* Beyond these traditional assumptions, the definition of $\mathcal{S}$ allows us to identify other cases that can guarantee linear convergence (e.g., in deterministic settings of Ω and various structures of $\boldsymbol{X}$ that does not satisfy incoherence property).

[5] In [33], C_R is some absolute constant that is independent of the problem parameters and the authors pick $\tau = 1/2$.

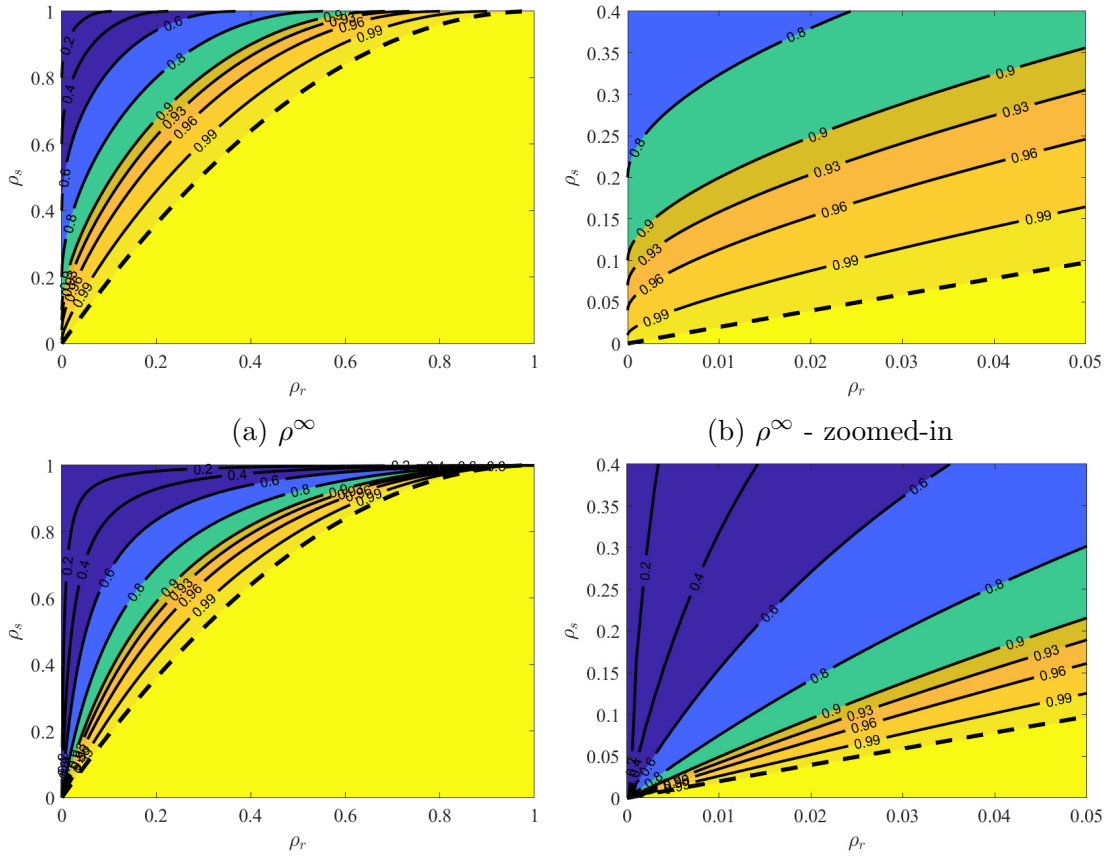
7.7.2 Convergence of IHT with the Optimal Step Size for Large-Scale Matrix Completion

It is noteworthy that the exact expression of the convergence rate provides more insights into the asymptotic behavior of PGD that can be **independent of the local structure** of the problem. As it is studied in Section IV of the original chapter, our result extends outside the fixed (low) rank regime considered in existing works and offer a way to evaluate the behavior of IHTSVD under more challenging conditions. In particular, when $\boldsymbol{U}$ and $\boldsymbol{V}$ are selected at random (e.g., from the Haar ensemble) with $r \sim O(\min\{n_1, n_2\})$ and $s \sim O(n_1 n_2)$, we show that the convergence rate approaches a limit which is independent of the actual matrix $\boldsymbol{X}$ and only depends on its dimensions (n_1, n_2), its rank r, and the sampling rate s:

$$\rho_\eta^\infty \to \max\Big\{|1-\eta\Big(\sqrt{(1-\rho_r)^2\rho_s}+\sqrt{\rho_r(2-\rho_r)(1-\rho_s)}\Big)^2|, \\ |1-\eta\Big(\sqrt{(1-\rho_r)^2\rho_s}-\sqrt{\rho_r(2-\rho_r)(1-\rho_s)}\Big)^2|\Big\}. \tag{7.24}$$

When $\eta = 1$, we have (7.24) becomes

$$\rho_1^\infty = 1-\Big(\sqrt{(1-\rho_r)^2\rho_s}-\sqrt{\rho_r(2-\rho_r)(1-\rho_s)}\Big)^2. \tag{7.25}$$



(a) ρ^{∞} (b) ρ^{∞} - zoomed-in

(c) ρ^{∞}_{opt} (d) ρ^{∞}_{opt} - zoomed-in

Figure 7.6: Contour plots of ρ_1^{∞} and ρ_{opt}^{∞} as 2-D functions of ρ_r and ρ_s given by (7.25) and (7.26), respectively. **(a) and (c):** the entire feasible range $\rho_r \in (0, 1]$ and $\rho_s \in (0, 1]$; **(b) and (d):** zoomed-in version of (a) and (c) near the bottom-left corner, respectively. The isoline at which $\rho_1^{\infty} = \rho_{opt}^{\infty} = 1$ is represented by the dashed line, corresponding to the case $\rho_s = 1 - (1 - \rho_r)^2$. The yellow region below this isoline corresponds to the under-determined setting $\rho_s < 1 - (1 - \rho_r)^2$. The common setting considered in the literature (e.g., [55, 104, 105] is the zoomed-in region where $\rho_r \ll \rho_s \ll 1$. On the other hand, our local convergence analysis covers the entire region in which the rank ratio and the sampling rate are not necessarily small.

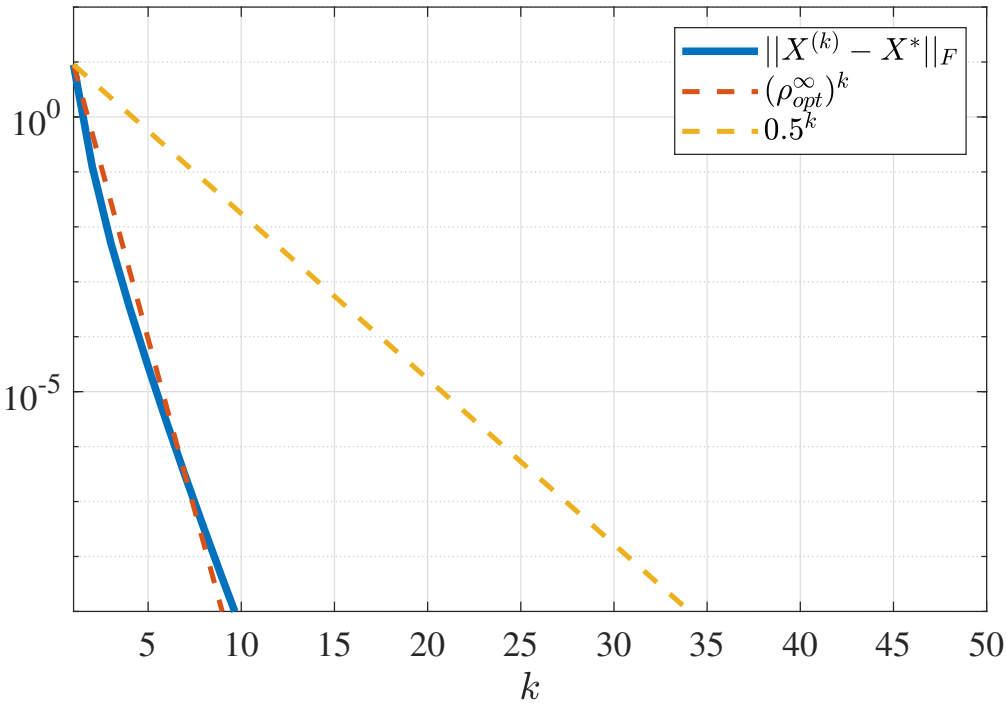


Figure 7.7: Convergence of IHT with step size $\eta = n_1 n_2/s$ under the setting $\rho_s = .2$ and $\rho_r = 0.0001$. With the matrix dimension being 10000, the difference between $\eta = n_1 n_2/s$ and the optimal step size η_{opt} given in (7.26) is as small as 0.003. The blue solid line represents the error through IHT iterations. The red and yellow dashed lines represent the exponential decrease at rates $\rho_{opt}^{\infty} = 0.056$ given in (7.26) and 0.5 given in [55], respectively. It can be seen that our estimate of the rate is tighter than the 0.5 global upper-bound in this asymptotic regime.

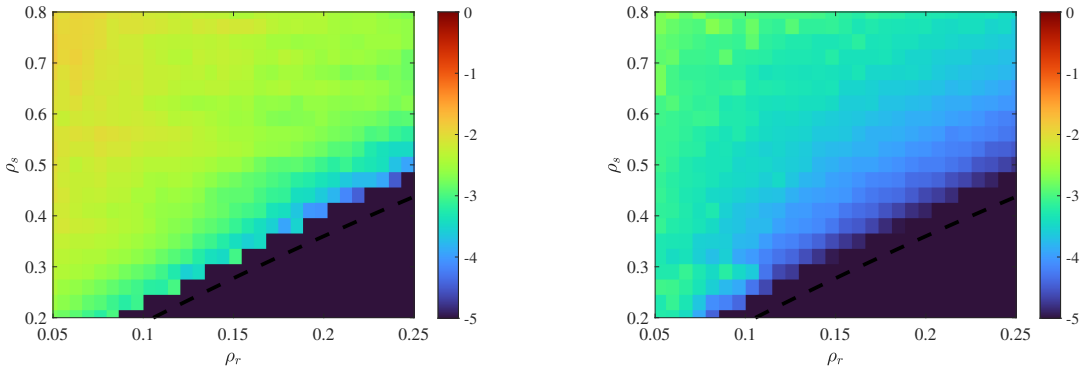
In addition, the optimal step size selected using this strategy and the corresponding optimal rate are given by

$$\eta_{opt}^{\infty} = \frac{1}{(1-\rho_r)^2\rho_s + \rho_r(2-\rho_r)(1-\rho_s)},$$
$$\rho_{opt}^{\infty} = \frac{2\sqrt{(1-\rho_r)^2\rho_s\rho_r(2-\rho_r)(1-\rho_s)}}{(1-\rho_r)^2\rho_s + \rho_r(2-\rho_r)(1-\rho_s)}. \tag{7.26}$$

Figure 7.6 demonstrates the rate of convergence in various setting of ρ_r and ρ_s. Note that if we evaluate this step-size choice under the regime suggested in [104, 105], i.e., $\lim_{\min\{n_1,n_2\}\to\infty} \rho_s = 0$ and $\lim_{m\to\infty} \rho_r/\rho_s^2 = 0$, then

$$\eta_{opt} = \frac{1}{\rho_s + 2\rho_r + o(\rho_s)} = \frac{1}{\rho_s}\Big(1 + o(\rho_s)\Big),$$
$$\rho_{opt} = 2\sqrt{\frac{2\rho_r}{\rho_s}}\big(1 + o(\rho_s)\big). \tag{7.27}$$

Comparing this with the step size $1/\rho_s$ selected in [55, 105], this provides the insight that the step size used in the approach of [105] not only guarantees linear convergence but also is optimal and cannot be improved upon. Notwithstanding, our local convergence analysis offers more precise estimate of the convergence rate compared to the 1/2 upper bound in prior works. In particular, in the aforementioned regime ($\rho_r \ll \rho_s$), our estimate of the rate ρ_{opt} approaches 0, which is much faster than the upper bound 1/2 (see Fig. 7.7).



(a) $n_1 = 500, n_2 = 400$ (b) $n_1 = 1200, n_2 = 1000$

Figure 7.8: The coefficient of variation (on a log10 scale) of the empirical rate shown in Fig. 7.5-(a) and (b), respectively. In each plot, the black dashed line corresponds to the boundary line $1 - \rho_s = (1 - \rho_r)^2$ and the black region on the bottom-right corner corresponds to the settings where no linear convergence is observed (i.e., the empirical rate is set to 1). The darker color in the right plot demonstrates the increasing concentration of the empirical rate as a random variable when the dimensions grow larger. It is also interesting to note that the variability in relation to the mean decreases as it approaches the boundary line (i.e., from the top-left corner to the bottom-right corner).

7.7.3 Proof of Theorem 7.1

7.7.3.1 Proof of Lemma 7.1

By the definition of the error matrix, we have

$$\begin{aligned}\boldsymbol{E}^{(k+1)} &= \boldsymbol{X}^{(k+1)} - \boldsymbol{M} \\ &= \Big(\mathcal{P}_{\bar{\Omega}}\big(\mathcal{P}_r(\boldsymbol{X}^{(k)})\big) + \mathcal{P}_{\Omega}(\boldsymbol{M})\Big) - \big(\mathcal{P}_{\Omega}(\boldsymbol{M}) + \mathcal{P}_{\bar{\Omega}}(\boldsymbol{M})\big) \\ &= \mathcal{P}_{\bar{\Omega}}\big(\mathcal{P}_r(\boldsymbol{M} + \boldsymbol{E}^{(k)}) - \boldsymbol{M}\big). \end{aligned} \tag{7.28}$$

From Proposition 7.1, we can reorganize (7.5) to obtain

$$\mathcal{P}_r(\boldsymbol{M} + \boldsymbol{E}^{(k)}) - \boldsymbol{M} = \boldsymbol{E}^{(k)} - \boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{E}^{(k)}\boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{E}^{(k)}).$$

Substituting the last equation back into (7.28) yields the recursion on the error matrix as in (7.12).

Next, let us denote $\boldsymbol{e}^{(k)} = \boldsymbol{S}_{\bar{\Omega}}^{\top}\mathrm{vec}(\boldsymbol{E}^{(k)})$, for $k = 1, 2, \ldots$. Vectorizing equation (7.12) and left-multiplying both sides with $\boldsymbol{S}_{\bar{\Omega}}$ yield

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{S}_{\bar{\Omega}}^{\top}\mathrm{vec}\Big(\mathcal{P}_{\bar{\Omega}}\big(\boldsymbol{E}^{(k)} - \boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{E}^{(k)}\boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{E}^{(k)})\big)\Big).$$

Using the property of selection matrices in Definition 7.2, we further have

$$\begin{aligned}\boldsymbol{e}^{(k+1)} &= \boldsymbol{S}_{\bar{\Omega}}^{\top}\boldsymbol{S}_{\bar{\Omega}}\boldsymbol{S}_{\bar{\Omega}}^{\top}\operatorname{vec}\big(\boldsymbol{E}^{(k)} - \boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{E}^{(k)}\boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{E}^{(k)})\big)\\ &= \boldsymbol{S}_{\bar{\Omega}}^{\top}\operatorname{vec}\big(\boldsymbol{E}^{(k)} - \boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{E}^{(k)}\boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{E}^{(k)})\big).\end{aligned}$$

Since $\operatorname{vec}(\boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{E}^{(k)}\boldsymbol{P}_{\boldsymbol{V}_\perp}) = (\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})\operatorname{vec}(\boldsymbol{E}^{(k)})$, the last equation can be represented as

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{S}_{\bar{\Omega}}^{\top}\operatorname{vec}(\boldsymbol{E}^{(k)}) - \boldsymbol{S}_{\bar{\Omega}}^{\top}(\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})\operatorname{vec}(\boldsymbol{E}^{(k)}) + \boldsymbol{S}_{\bar{\Omega}}^{\top}\operatorname{vec}\big(\boldsymbol{R}(\boldsymbol{E}^{(k)})\big). \quad (7.29)$$

On the other hand, (7.12) implies, for any $k \geq 1$, $\boldsymbol{E}^{(k)} = \mathcal{P}_{\bar{\Omega}}(\boldsymbol{E}^{(k)})$ and

$$\operatorname{vec}(\boldsymbol{E}^{(k)}) = \operatorname{vec}\big(\mathcal{P}_{\bar{\Omega}}(\boldsymbol{E}^{(k)})\big) = \boldsymbol{S}_{\bar{\Omega}}\boldsymbol{S}_{\bar{\Omega}}^{\top}\operatorname{vec}(\boldsymbol{E}^{(k)}) = \boldsymbol{S}_{\bar{\Omega}}\boldsymbol{e}^{(k)}.$$

Substituting the last equation into the RHS of (7.29) yields (7.13).

7.7.3.2 Proof of Lemma 7.2

Applying the triangle inequality to the RHS of (7.13) yields

$$\left\|\boldsymbol{e}^{(k+1)}\right\|_2 \leq \left\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{e}^{(k)}\right\|_2 + \left\|\boldsymbol{r}(\boldsymbol{e}^{(k)})\right\|_2, \quad (7.30)$$

where we recall $\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^{\top}(\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})\boldsymbol{S}_{\bar{\Omega}}$. By the definition of the operator norm, we have

$$\begin{aligned}\left\|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{e}^{(k)}\right\|_2 &\leq \|\boldsymbol{I}-\boldsymbol{H}\|_2 \left\|\boldsymbol{e}^{(k)}\right\|_2 \\ &= \max_i\left\{|1-\lambda_i(\boldsymbol{H})|\right\} \cdot \left\|\boldsymbol{e}^{(k)}\right\|_2 \\ &= \left(1-\lambda_{\min}(\boldsymbol{H})\right)\left\|\boldsymbol{e}^{(k)}\right\|_2, \end{aligned} \tag{7.31}$$

where the last equality stems from the fact that all eigenvalues of $\boldsymbol{H}$ lie between 0 and 1. From (7.30) and (7.31), we obtain

$$\left\|\boldsymbol{e}^{(k+1)}\right\|_2 \leq \left(1-\lambda_{\min}(\boldsymbol{H})\right)\left\|\boldsymbol{e}^{(k)}\right\|_2 + \left\|\boldsymbol{r}(\boldsymbol{e}^{(k)})\right\|_2. \tag{7.32}$$

The conclusion of lemma follows from the fact that

$$\left\|\boldsymbol{e}^{(k)}\right\|_2 = \left\|\mathcal{P}_{\bar{\Omega}}\left(\boldsymbol{E}^{(k)}\right)\right\|_F = \left\|\boldsymbol{E}^{(k)}\right\|_F$$

and

$$\left\|\boldsymbol{r}(\boldsymbol{e}^{(k)})\right\|_2 \leq \left\|\boldsymbol{R}(\boldsymbol{E}^{(k)})\right\|_F \leq \frac{c_1}{\sigma_r}\left\|\boldsymbol{E}^{(k)}\right\|_F^2.$$

7.7.4 Details of Example 7.1

7.7.4.1 The first case

Using the same argument as in Lemma 5.3 in [232], we can replace the complex matrix in (7.21) by a real PSD matrix and prove the following lemma:

Lemma 7.5. *Let $\boldsymbol{a} = [a_1, \ldots, a_{qn}]^\top$ is a random vector with i.i.d entries, where $a_i \sim \mathcal{N}(0, 1/n)$. Then for any sequence of $qn \times qn$ PSD matrices $\boldsymbol{M}_{qn}$ with uniformly bounded spectral norms $\|\boldsymbol{M}_{qn}\|_2$, we have*

$$\left(\boldsymbol{a}^\top \boldsymbol{M}_{qn} \boldsymbol{a} - \frac{1}{n} \operatorname{tr}(\boldsymbol{M}_{qn})\right) \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Proof. To simplify our notation, let us denote the (i, j)-th entry of $\boldsymbol{M}_{qn}$ by M_{ij} and δ_{ij} is the indicator of the event $i = j$. Since a_i are $i.i.d$ normally distributed, we have

$$\mathbb{E}[a_i] = 0, \quad \mathbb{E}[a_i a_j] = \delta_{ij} \frac{1}{n}, \qquad \mathbb{E}[a_i a_j a_k a_l] = (\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) \frac{1}{n^2}, \quad (7.33)$$

for any indices $1 \leq i, j, k, l \leq n$. In order to prove $\left(\boldsymbol{a}^\top \boldsymbol{M}_{qn} \boldsymbol{a} - \frac{1}{n} \operatorname{tr}(\boldsymbol{M}_{qn})\right) \xrightarrow{\mathrm{p}} 0$, it is sufficient to show that

$$\begin{cases} \mathbb{E}[\boldsymbol{a}^\top \boldsymbol{M}_{qn} \boldsymbol{a}] = \frac{1}{n} \operatorname{tr}(\boldsymbol{M}_{qn}), \\ \operatorname{Var}(\boldsymbol{a}^\top \boldsymbol{M}_{qn} \boldsymbol{a}) \to 0 \text{ as } n \to \infty. \end{cases}$$

First, by the linearity of expectation, we have

$$\begin{aligned}\mathbb{E}[\boldsymbol{a}^\top \boldsymbol{M}_{qn}\boldsymbol{a}] &= \mathbb{E}\Big[\sum_{i,j} M_{ij}a_i a_j\Big] = \sum_{i,j} M_{ij}\mathbb{E}[a_i a_j] \\ &= \sum_{i,j} M_{ij}\delta_{ij}\frac{1}{n} = \frac{1}{n}\sum_{i=1}^{qn} M_{ii} = \frac{1}{n}\operatorname{tr}(\boldsymbol{M}_{qn}). \end{aligned} \tag{7.34}$$

Second, by rewriting the variance of the summation $\sum_{i,j} M_{ij}a_i a_j$ in terms of the sum of covariances, we obtain

$$\begin{aligned}\operatorname{Var}(\boldsymbol{a}^\top \boldsymbol{M}_{qn}\boldsymbol{a}) &= \operatorname{Var}\Big(\sum_{i,j} M_{ij}a_i a_j\Big) \\ &= \sum_{i,j,k,l} \operatorname{Cov}(M_{ij}a_i a_j, M_{kl}a_k a_l). \end{aligned} \tag{7.35}$$

Using the formula

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \tag{7.36}$$

and the linearity of expectation, (7.35) can be represented as

$$\begin{aligned}\operatorname{Var}(\boldsymbol{a}^\top \boldsymbol{M}_{qn}\boldsymbol{a}) &= \sum_{i,j,k,l} M_{ij}M_{kl}\Big(\mathbb{E}[a_i a_j a_k a_l] - \mathbb{E}[a_i a_j]\mathbb{E}[a_k a_l]\Big) \\ &= \sum_{i,j,k,l} M_{ij}M_{kl}\big(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\big)\frac{1}{n^2} \\ &= \frac{2}{n^2}\sum_{i,j} M_{ij}^2 = \frac{2}{n^2}\left\|\boldsymbol{M}_{qn}\right\|_F^2. \end{aligned} \tag{7.37}$$

Since $\boldsymbol{M}_{qn}$ is PSD and has bounded spectral norm, all of its eigenvalues are

bounded by $0 \leq \lambda_i(\boldsymbol{M}_{qn}) \leq C$, for some constant C, and hence,

$$\|\boldsymbol{M}_{qn}\|_F^2 = \sum_{i=1}^{qn} \lambda_i^2(\boldsymbol{M}_{qn}) \leq qnC^2.$$

Thus, substituting back into (7.37) yields

$$\mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{M}_{qn} \boldsymbol{a}) \leq \frac{2}{n^2} qnC^2 \to 0 \quad \text{as} \quad n \to \infty.$$

This completes our proof of the lemma. □

7.7.4.2 The second case

Similarly, we consider the following lemma:

Lemma 7.6. *Let $\boldsymbol{b} = [b_1, \ldots, b_{qn}]$ and $\boldsymbol{c} = [c_1, \ldots, c_{qn}]$ are random vectors with i.i.d entries, where $b_i, c_j \sim \mathcal{N}(0, 1/n)$. Denote $m = n^2$, $k = q^2$ and $\boldsymbol{a} = \boldsymbol{b} \otimes \boldsymbol{c}$. Then for any sequence of $km \times km$ PSD matrices $\boldsymbol{M}_{km}$ with uniformly bounded spectral norms $\|\boldsymbol{M}_{km}\|_2$, we have*

$$\left(\boldsymbol{a}^\top \boldsymbol{M}_{km} \boldsymbol{a} - \frac{1}{m} \mathrm{tr}(\boldsymbol{M}_{km})\right) \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Proof. Denote $\boldsymbol{M}_{[ij]}$ is the (i, j)-th $qn \times qn$ block of $\boldsymbol{M}_{km}$. Then it is straightforward

to verify that

$$\boldsymbol{a}^\top \boldsymbol{M}_{km} \boldsymbol{a} = \sum_{i,j} b_i (\boldsymbol{c}^\top \boldsymbol{M}_{[ij]} \boldsymbol{c}) b_j.$$

In order to prove $\big(\boldsymbol{a}^\top \boldsymbol{M}_{km} \boldsymbol{a} - \frac{1}{m} \operatorname{tr}(\boldsymbol{M}_{km})\big) \xrightarrow{\text{p}} 0$, it is sufficient to show that

$$\begin{cases} \mathbb{E}[\boldsymbol{a}^\top \boldsymbol{M}_{km} \boldsymbol{a}] = \frac{1}{m} \operatorname{tr}(\boldsymbol{M}_{km}), \\ \operatorname{Var}(\boldsymbol{a}^\top \boldsymbol{M}_{km} \boldsymbol{a}) \to 0 \text{ as } n \to \infty. \end{cases}$$

First, we use the linearity of expectation to obtain

$$\begin{aligned} \mathbb{E}[\boldsymbol{a}^\top \boldsymbol{M}_{km} \boldsymbol{a}] &= \mathbb{E}\Big[\sum_{i,j} b_i (\boldsymbol{c}^\top \boldsymbol{M}_{[ij]} \boldsymbol{c}) b_j\Big] \\ &= \sum_{i,j} \mathbb{E}[b_i b_j] \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{M}_{[ij]} \boldsymbol{c}]. \end{aligned}$$

From (7.34) and Lemma 7.5, the last equation is equivalent to

$$\mathbb{E}[\boldsymbol{a}^\top \boldsymbol{M}_{km} \boldsymbol{a}] = \sum_{i,j} \delta_{ij} \frac{1}{n} \cdot \frac{1}{n} \operatorname{tr}(\boldsymbol{M}_{[ij]}) = \frac{1}{m} \operatorname{tr}(\boldsymbol{M}_{km}).$$

Second, we have

$$\begin{aligned} \operatorname{Var}(\boldsymbol{a}^\top \boldsymbol{M}_{km} \boldsymbol{a}) &= \operatorname{Var}\big(\sum_{i,j} b_i (\boldsymbol{c}^\top \boldsymbol{M}_{[ij]} \boldsymbol{c}) b_j\big) \\ &= \sum_{i,j,k,l} \operatorname{Cov}\big(b_i (\boldsymbol{c}^\top \boldsymbol{M}_{[ij]} \boldsymbol{c}) b_j, b_k (\boldsymbol{c}^\top \boldsymbol{M}_{[kl]} \boldsymbol{c}) b_l\big). \end{aligned} \tag{7.38}$$

From (7.36), each covariance on the RHS of (7.38) can be represented as

$$\begin{aligned}\mathrm{Cov}\big(b_i(\boldsymbol{c}^\top \boldsymbol{M}_{[ij]}\boldsymbol{c})b_j, b_k(\boldsymbol{c}^\top \boldsymbol{M}_{[kl]}\boldsymbol{c})b_l\big) &= \mathbb{E}[b_i b_j b_k b_l]\cdot \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{M}_{[ij]}\boldsymbol{c}\cdot \boldsymbol{c}^\top \boldsymbol{M}_{[kl]}\boldsymbol{c}] \\ &\quad - \mathbb{E}[b_i b_j]\cdot \mathbb{E}[b_k b_l]\cdot \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{M}_{[ij]}\boldsymbol{c}]\cdot \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{M}_{[kl]}\boldsymbol{c}].\end{aligned} \tag{7.39}$$

Lemma 7.7. *Let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be matrices in $\mathbb{R}^{qn\times qn}$. Then*

$$\mathbb{E}[\boldsymbol{c}^\top \boldsymbol{P}\boldsymbol{c}\cdot \boldsymbol{c}^\top \boldsymbol{Q}\boldsymbol{c}] = \frac{\mathrm{tr}(\boldsymbol{P})\,\mathrm{tr}(\boldsymbol{Q}) + \mathrm{tr}(\boldsymbol{P}\boldsymbol{Q}^\top) + \mathrm{tr}(\boldsymbol{P}\boldsymbol{Q})}{n^2}.$$

The proof of Lemma 7.7 is straightforward from (7.33) and is omitted in this chapter. From Lemma 7.7 and (7.33), we can simplify (7.39) as

$$\begin{aligned}\mathrm{Cov}\big(b_i(\boldsymbol{c}^\top \boldsymbol{M}_{[ij]}\boldsymbol{c})b_j, b_k(\boldsymbol{c}^\top \boldsymbol{M}_{[kl]}\boldsymbol{c})b_l\big) = \frac{1}{n^4}\Big(&\mathrm{tr}(\boldsymbol{M}_{[ij]}\boldsymbol{M}_{[kl]}) + \mathrm{tr}(\boldsymbol{M}_{[ij]}\boldsymbol{M}_{[kl]}^\top) \\ &+ \mathrm{tr}^2(\boldsymbol{M}_{[ij]}) + \mathrm{tr}(\boldsymbol{M}_{[ij]}^2) + \mathrm{tr}(\boldsymbol{M}_{[ij]}\boldsymbol{M}_{[ij]}^\top) \\ &+ \mathrm{tr}(\boldsymbol{M}_{[ij]})\,\mathrm{tr}(\boldsymbol{M}_{[ij]}^\top) + \mathrm{tr}(\boldsymbol{M}_{[ij]}^2)\Big).\end{aligned}$$

Substituting the last equation back into (7.38) yields

$$\begin{aligned}\mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{M}_{km}\boldsymbol{a}) = \frac{2}{n^4}\Big(&\sum_{i,j}\mathrm{tr}^2(\boldsymbol{M}_{[ij]}) + \sum_{i,j}\mathrm{tr}(\boldsymbol{M}_{[ii]}\boldsymbol{M}_{[jj]}) \\ &+ \sum_{i,j}\mathrm{tr}(\boldsymbol{M}_{[ij]}^\top \boldsymbol{M}_{[jj]}) + \sum_{i,j}\mathrm{tr}(\boldsymbol{M}_{[ij]}^2)\Big).\end{aligned} \tag{7.40}$$

Next, we bound each term on the RHS of (7.40). To that end, we utilize the

following lemma:

Lemma 7.8. *For any matrices* $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$*, it holds that*

1. $\|\boldsymbol{A}\|_F \leq \sqrt{n}\,\|\boldsymbol{A}\|_2$,
2. $\mathrm{tr}^2(\boldsymbol{A}) \leq n\,\|\boldsymbol{A}\|_F^2$,
3. $\mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{B}) \leq \|\boldsymbol{A}\|_F\,\|\boldsymbol{B}\|_F \leq n\,\|\boldsymbol{A}\|_2\,\|\boldsymbol{B}\|_2$,
4. $\mathrm{tr}(\boldsymbol{A}^2) \leq \|\boldsymbol{A}\|_F^2 = \mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{A})$.

The proof of Lemma 7.8 can be found in [151] - Chapter 5. Applying Lemma 7.8 with the blocks of size $qn \times qn$, we obtain

$$\begin{aligned}\sum_{i,j}\mathrm{tr}^2(\boldsymbol{M}_{[ij]}) &\leq \sum_{i,j} qn\,\left\|\boldsymbol{M}_{[ij]}\right\|_F^2 = qn\,\|\boldsymbol{M}\|_F^2\\ &\leq (qn)^3\,\|\boldsymbol{M}\|_2 \leq C(qn)^3,\end{aligned}$$

$$\begin{aligned}\sum_{i,j}\mathrm{tr}(\boldsymbol{M}_{[ii]}\boldsymbol{M}_{[jj]}) &\leq \sum_{i,j} qn\,\left\|\boldsymbol{M}_{[ii]}\right\|_2\left\|\boldsymbol{M}_{[jj]}\right\|_2\\ &\leq \sum_{i,j} qn\,\|\boldsymbol{M}\|_2\,\|\boldsymbol{M}\|_2 = C^2(qn)^3,\end{aligned}$$

$$\sum_{i,j}\mathrm{tr}(\boldsymbol{M}_{[ij]}^\top\boldsymbol{M}_{[jj]}) = \sum_{i,j}\left\|\boldsymbol{M}_{[ij]}\right\|_F^2 = \|\boldsymbol{M}\|_F^2 \leq C(qn)^2,$$

$$\sum_{i,j} \operatorname{tr}(\boldsymbol{M}_{[ij]}^2) \leq \sum_{i,j} \left\|\boldsymbol{M}_{[ij]}\right\|_F^2 = \|\boldsymbol{M}\|_F^2 \leq C(qn)^2.$$

Therefore, (7.40) can be bounded as

$$\operatorname{Var}(\boldsymbol{a}^\top \boldsymbol{M}_{km} \boldsymbol{a}) \leq \frac{2}{n^4}(C(qn)^3 + C^2(qn)^3 + 2C(qn)^2).$$

The conclusion of the lemma follows by the fact that the RHS of the last equation which approaches 0 as $n \to \infty$. □

Chapter 8: Accelerating Iterative Hard Thresholding for Low-Rank Matrix Completion via Adaptive Restart[1]

This chapter introduces the use of adaptive restart to accelerate iterative hard thresholding (IHT) for low-rank matrix completion. First, we analyze the local convergence of accelerated IHT in the non-convex setting of matrix completion problem (MCP). We prove the linear convergence rate of the accelerated algorithm inside the region near the solution. Our analysis poses a major challenge to parameter selection for accelerated IHT when no prior knowledge of the "local Hessian condition number" is given. To address this issue, we propose a simple adaptive restart algorithm for MCP to recover the optimal rate of convergence at the solution, as motivated in [164]. Our numerical result verifies the theoretical analysis as well as demonstrates the outstanding performance of the proposed algorithm.

[1]This work has been published as: Trung Vu and Raviv Raich. "Accelerating Iterative Hard Thresholding for Low-Rank Matrix Completion via Adaptive Restart." In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2917-2921. IEEE, 2019.

8.1 Introduction

Low-rank matrix completion is a fundamental problem that arises in many areas of signal processing and machine learning such as collaborative filtering [176], system identification [137] and dimension reduction [31]. The problem can be explained as follows. Let $M \in \mathbb{R}^{m\times n}$ be the underlying matrix with low rank r and a subset its entries $\mathcal{S} = \{(i,j) \mid M_{ij} \text{ is observed}\}$. We aim to recover the unknown entries of M, belonging to the complement set $\mathcal{S}^c$. Alternatively, one would solve the following optimization problem:

$$\min_{X\in\mathbb{R}^{m\times n}} \operatorname{rank}(X) \quad \text{s.t. } X_{ij} = M_{ij}, \quad \forall (i,j) \in \mathcal{S}. \tag{8.1}$$

In one of the pioneer works, Candès and Recht [33] introduced a convex relaxation to the original non-convex matrix completion problem and presented conditions under which the solutions of the two problems coincide. Moreover, they provided an expression for the number of known entries required to recover the original matrix. This breakthrough leads to the class of proximal-type algorithms for nuclear norm minimization [27, 107, 145, 205] with rigorous mathematical guarantees and extensions of classic acceleration techniques. The disadvantage of convex-relaxed methods, nonetheless, is either high computational complexity (for interior-point methods) or slow convergence rate (often sublinear for proximal-type methods). To address those issues, iterative hard thresholding has been proposed to directly solve the non-convex rank minimization problem [79,104,125]. Each IHT iteration takes one step in the direction of the gradient and one step projecting onto the

set of rank-r matrices. Since the process is akin to hard-thresholding singular values, we refer to the methods using it as iterative hard thresholding algorithms, as opposed to their aforementioned soft thresholding counterparts. When the solution is low-rank, IHT is extremely efficient in both computational complexity and empirical convergence (linear rate). Notwithstanding, mathematical guarantees of non-convex IHT algorithms for MCP are generally restricted to local convergence [46, 120].

Despite the similarity between IHT and projected gradient methods, there have been but a few efforts in accelerating IHT and characterizing the performance thereby. In a very recent work, Khanna and Kyrillidis [116] introduces the use of acceleration to plain IHT yet in the context of rank minimization with affine constraints (ARMP). The authors provided convergence guarantees based on restricted strong convexity and smoothness properties of the loss function. However, as pointed out in [33], the results and techniques for ARMP cannot apply to MCP for which the restricted properties does not hold. Additionally, they left an open question on the optimal momentum step sizes that guarantee better performance over plain IHT. While determining an optimal tuning is NP-hard [204], our experiment indeed shows that a careless choice of step sizes might worsen the performance of plain IHT in a matrix completion setting. Thus, we believe answering this question is the key to the practicality of accelerated IHT in both ARMP and MCP.

In this chapter, we consider IHT for solving low-rank matrix completion and connect the classic theory of accelerated gradient methods with recent analyses of

the local convergence of plain IHT in [46]. The contribution of our work is threefold: (i) we propose a variant of Nesterov's Accelerated Gradient in a MCP-IHT setting and analyze the local convergence thereof, (ii) we identify the choice of momentum step sizes that guarantees the optimal acceleration, (iii) we propose an adaptive restart algorithm that can asymptotically recover the local rate in practice. The numerical experiment verifies our theoretical analysis and demonstrates the superior performance of the proposed algorithm compared to common existing methods for low-rank matrix completion.

8.2 Preliminaries

We begin with a review of some preliminaries on iterative hard thresholding methods for low-rank matrix completion.

Definition 8.1. *Let $M \in \mathbb{R}^{m\times n}(m \geq n)$ be a rank-r matrix and $M = U\Sigma V^T$ be its singular value decomposition (SVD), where Σ is a diagonal $m \times n$ matrix with diagonal entries*

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$$

and U, V are $m \times m$ and $n \times n$ unitary matrices, respectively. We partition U, Σ, V

as follows:

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$$

where $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, $\Sigma_2 = \mathbf{0}$*;* U_1, V_1 *and* U_2, V_2 *are semi-unitary matrices corresponding to the partition of* Σ*.*

Definition 8.2. *A row selection matrix* $S \in \mathbb{R}^{s \times m} (s \leq m)$ *is a semi-unitary matrix obtained by a subset of* s *rows from the identity matrix* I_m*. Left-multiplying a matrix* $X \in \mathbb{R}^{m \times n}$ *by* S *returns an* $s \times n$ *matrix corresponding to set of rows in* X*.*

Definition 8.3. *Sampling operator* $X_{\mathcal{S}}$ *maps the matrix entries not in* $\mathcal{S}$ *to* 0*:*

$$[X_{\mathcal{S}}]_{ij} = \begin{cases} X_{ij} & \text{if } (i,j) \in \mathcal{S}, \\ 0 & \text{if } (i,j) \in \mathcal{S}^c. \end{cases}$$

Definition 8.4. *Let* $X \in \mathbb{R}^{m \times n}$ *be a matrix with arbitrary rank. Define the rank-r projection of* X *as:*

$$\mathcal{P}_r(X) \in \underset{Y \in \mathbb{R}^{m \times n}}{\mathrm{argmin}} \|Y - X\|_F \;\; s.t. \;\; \mathrm{rank}(Y) \leq r.$$

The solution of this minimization is obtained by computing the top r singular values and vectors of X [61]. Moreover, this projection is unique if either $\sigma_r(X) > \sigma_{r+1}(X)$ or $\sigma_r(X) = 0$, where $\sigma_r(.)$ denotes the r-th largest singular value. In

the rest of this chapter, we implicitly refer the solution of problem (8.1) and its SVD to the notations in Definition 8.1. This also implies our assumption that $\text{rank}(M) = r$. Furthermore, we denote the cardinality of $\mathcal{S}$ by s and the row selection matrix corresponding to $\mathcal{S}^c$ by $S_c \in \mathbb{R}^{(mn-s)\times mn}$.

8.3 Background

8.3.1 ARMP-IHT versus MCP-IHT

Iterative hard thresholding for low-rank matrices was first proposed in the context of ARMP. In [104], Jain et. al. considered the following robust formulation of ARMP:

$$\min_{X\in\mathbb{R}^{m\times n}} \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2 \text{ s.t. } \text{rank}(X) \leq r \tag{8.2}$$

where $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^s$ is an affine transformation and $b \in \mathbb{R}^s$ is the set of indirect observations. Adapting the idea of projected gradient descent, the authors proposed the Singular Value Projection (SVP) algorithm with the basic update

$$X^{(k)} = \mathcal{P}_r(X^{(k-1)} - \eta_k \mathcal{A}^T(\mathcal{A}(X^{(k-1)}) - b)).$$

Under assumptions on Restricted Isometry Property (RIP) of the affine operator $\mathcal{A}$, the authors showed that their algorithm converges to the solution at a linear rate. In an independent work, Goldfarb and Ma [79] proved the geometric convergence

Algorithm 8.1 Iterative Hard Thresholding

1: $X^{(0)} = M_{\mathcal{S}}$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad X^{(k)} = \mathcal{P}_r\big(X^{(k-1)} - \alpha_k[X^{(k-1)} - M]_{\mathcal{S}}\big)$

for a special case of unit step size. Later on, there have been efforts in improving the performance of ARMP-IHT, namely Normalized IHT [201] and accelerated IHT [116]. All of these works use the standard RIP assumptions in order to prove the global convergence.

The matrix completion problem is a special case of ARMP where the affine operator $\mathcal{A}$ is a sampling operator:

$$\min_{X \in \mathbb{R}^{m \times n}} \|X_{\mathcal{S}} - M_{\mathcal{S}}\|_F^2 \text{ s.t. } \operatorname{rank}(X) \leq r. \tag{8.3}$$

Unfortunately, this operator does not satisfy RIP in general, shattering the global convergence guarantees established in ARMP. Still, Jain et. al. suggested to apply SVP for solving MCP (see Algorithm 8.1) and made a conjecture that SVP converges linearly to the solution matrix M with high probability, provided M is incoherent [104]. It took some time before the first theoretical guarantee is obtained in [46], considering a special case of SVP, called IHTSVD algorithm. When the step size α_k equals 1, one can simplify the gradient update in Algorithm 8.1 as $X^{(k-1)} - \alpha_k[X^{(k-1)} - M]_{\mathcal{S}} = [X^{(k-1)}]_{\mathcal{S}^c} + M_{\mathcal{S}}$. For convenience, we call this operator the *observation projection*, denoted by $\mathcal{P}_{M,\mathcal{S}}$. It simply sets entries of $X^{(k)}$ that are in $\mathcal{S}$ to those corresponding values of M. The IHT iterates now serve as alternating projections between $\mathcal{P}_r$ and $\mathcal{P}_{M,\mathcal{S}}$. More importantly, the authors

provided a quantitative analysis on the local convergence of IHTSVD, based on the approximation of rank-r projection operator near the solution. Let us restate their results in Theorem 8.1 and Theorem 8.2. We use our own notations for the purpose of consistency.

Theorem 8.1. *(Rephrased from [46]) Given the matrix M in Definition 8.1. Denote $\epsilon = \min_{\sigma_i > \sigma_{i+1}} \{\sigma_i - \sigma_{i+1}\}$. Let $\Delta \in \mathbb{R}^{m\times n}$ be a perturbation matrix such that $\|\Delta\|_F < \frac{\epsilon}{2}$. Then the rank-$r$ projection of $M + \Delta$ is given by*

$$\mathcal{P}_r(M + \Delta) = M + \Delta - U_2 U_2^T \Delta V_2 V_2^T + Q(\Delta)$$

where $Q : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ satisfies $\|Q(\Delta)\|_F = O(\|\Delta\|_F^2)$.

Theorem 8.2. *(Rephrased from [46]) If the matrix $S_c(V_2 \otimes U_2)$ has full rank, then Algorithm 8.1 with a unit step size converges to M locally at a linear rate $1 - \sigma^2$, where $\sigma = \sigma_{\min}\big(S_c(V_2 \otimes U_2)\big)$. In other words, there exists a neighborhood $\mathcal{E}(M)$ of M and a constant C such that if $X^{(0)} \in \mathcal{E}(M)$, then*

$$\big\|X^{(k)} - M\big\|_F \leq C\big(1 - \sigma^2\big)^k \big\|X^{(0)} - M\big\|_F .$$

8.3.2 Nesterov's Accelerated Gradient for ARMP-IHT

We consider the plain IHT as a first-order gradient method and apply momentum techniques to accelerate it. In [160], Nesterov demonstrated a simple modification to gradient descent that provably improves the convergence rate dramatically. The

method, known as Nesterov's Accelerated Gradient (NAG), can be described as follows

$$x^{(k)} = y^{(k-1)} - \alpha_k \nabla f(y^{(k-1)})$$
$$y^{(k)} = x^{(k)} + \beta_k (x^{(k)} - x^{(k-1)})$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous differentiable, smooth convex function to be minimized. For optimizing an μ-strongly convex, L-smooth function, it is well-known that NAG obtains a linear convergence rate at $1 - \sqrt{\mu/L}$ by setting [160]

$$\alpha_k = \frac{1}{L}, \qquad \beta_k = \frac{1 - \sqrt{\mu/L}}{1 + \sqrt{\mu/L}}. \tag{8.4}$$

This scheme is often called optimal since it achieves the lower complexity bound for first-order methods on minimizing a strongly convex, smooth function derived by Nemirovski and Yudin [158].

The idea of accelerating IHT has recently been studied in [116,119] for ARMP. It is similar in spirit to the Accelerated Proximal Gradient algorithm for solving nuclear norm regularized linear least square problems [205]. While these algorithms enjoy the convexity of the norm operator and copious theoretical guarantees of proximal methods, the burden of non-convex projections over the rank constraint bears heavily on IHT methods. Moreover, as we mentioned, convergence guarantee for accelerated ARMP-IHT in [116] does not hold for MCP-IHT. To the best of our knowledge, there is no convergence analysis for accelerated IHT in a matrix

completion setting to date.

8.4 Accelerating MCP-IHT

In this section, we first describe an accelerated scheme for Algorithm 8.1 and provide some analysis of the local convergence of the algorithm. It remains a challenging problem on parameter selection that guarantees better performance of accelerated over plain IHT. To address this issue, we propose an adaptive restart technique that allows us to asymptotically recover the optimal rate of convergence in practice.

8.4.1 An NAG-variant of MCP-IHT

Motivated by the result in Theorem 8.2, we propose an NAG-variant of IHT in Algorithm 8.2. First, notice the specific choice of gradient step size ($\alpha_k = 1$) unveils the observation projection $\mathcal{P}_{M,\mathcal{S}}$. Interestingly, this choice of α_k matches the setting in (8.4), as the Lipschitz constant of the sampling operator is $L = 1$. Second, the order at each iteration guarantees the sequence $\{Y^{(k)}\}$ is consistent with the observation $\mathcal{S}$, i.e., $Y^{(k)}_{\mathcal{S}} = M_{\mathcal{S}}$. As a result, the error matrix depends only on the entries in $\mathcal{S}^c$, disentangling the subsequent analysis of convergence. Finally, the algorithm terminates when a stopping criteria is met, returning $Y^{(k)}$ as an estimate of the matrix. We state our main theoretical result for the convergence of

Algorithm 8.2 NAG-IHT

1: $X^{(0)} = Y^{(0)} = M_{\mathcal{S}}$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad X^{(k)} = \mathcal{P}_r\big(Y^{(k-1)}\big)$
4: $\quad Y^{(k)} = \mathcal{P}_{M,\mathcal{S}}\big(X^{(k)} + \beta_k(X^{(k)} - X^{(k-1)})\big)$

NAG-IHT in Theorem 8.3.[2] Note that the convergence rate is described in a closed-form, which can be verified through experiments. By contrast, RIP constants in the standard analysis for ARMP are NP-Hard to compute [204]. Mainly, the optimal fixed step size for NAG-IHT is identified, guaranteeing the better performance of accelerated schemes over plain IHT in theory, i.e., $1-\sigma$ versus $1-\sigma^2$.

Theorem 8.3. *If the matrix $S_c(V_2 \otimes U_2)$ has full rank, then Algorithm 8.2 with momentum step size $\beta_k = \frac{1-\sigma}{1+\sigma}$ converges to M locally at a linear rate $1-\sigma$, where $\sigma = \sigma_{\min}(S_c(V_2 \otimes U_2))$. In other words, there exists a neighborhood $\mathcal{E}(M)$ of M and a constant C such that if $Y^{(0)} \in \mathcal{E}(M)$, then*

$$\big\|Y^{(k)} - M\big\|_F \leq C\big(1-\sigma\big)^k \big\|Y^{(0)} - M\big\|_F .$$

Further, this is the optimal rate for any fixed momentum step size in Algorithm 8.2.

[2]The proofs of Theorem 8.1 and Theorem 8.3 are given in the Appendix at the end of this chapter.

Algorithm 8.3 ARNAG-IHT

$t = 1, X^{(0)} = Y^{(0)} = M_{\mathcal{S}}, f_0 = \left\| X_{\mathcal{S}}^{(0)} - M_{\mathcal{S}} \right\|_F^2$
for $k = 1, 2, \ldots$ **do**
  $X^{(k)} = \mathcal{P}_r\big(Y^{(k-1)}\big)$
  $Y^{(k)} = \mathcal{P}_{M,\mathcal{S}}\big(X^{(k)} + \frac{t-1}{t+2}(X^{(k)} - X^{(k-1)})\big)$
  $f_k = \left\| X_{\mathcal{S}}^{(k)} - M_{\mathcal{S}} \right\|_F^2$
  if $f_k > f_{k-1}$ **then** $t = 1$ **else** $t = t + 1$

8.4.2 An Adaptive Restart Scheme for NAG-IHT

Theorem 8.3 provides a theoretical guarantee for NAG-IHT but it implies that fixed-step-size strategy is impracticable when the value of σ is unknown. In this section, we propose a simple way to work around this issue. The idea stems from adaptive restart techniques for accelerated gradient schemes [164]: reset the momentum back to zero whenever we observe an increase in the function value. This facile heuristic was shown to asymptotically recover the local rate of convergence of NAG on minimizing a strongly convex smooth function and is generally used in sparse signal recovery. To the best of our knowledge, this work is the first to adopt adaptive restart heuristics to accelerate IHT. We describe our approach, named ARNAG-IHT, in Algorithm 8.3. It is important to highlight that the momentum need to grows from one iteration to the next in order to apply restart techniques. As a result, we use the incremental step size $\beta_k = \frac{t-1}{t+2}$ recommended in optimizing smooth convex functions [160]. The difference comes with conditional restarts (setting $t = 1$) whenever the square loss increases. Clearly, all three aforementioned algorithms share the same computational complexity per iteration.

8.5 Empirical Result

This section presents a numerical example to demonstrate our analysis for low-rank matrix completion. First, we generate a solution matrix $M \in \mathbb{R}^{m\times n}$ of rank r by taking the product of an $m \times r$ matrix and an $r \times n$ matrix, each having i.i.d. normally distributed entries. Next, we sample the observation set $\mathcal{S}$ uniformly at random. We compare ARNAG-IHT with the following methods: SVT [27], SVP-NewtonD [104], NIHT [201] and IHTSVD [46]. Although the analyses of SVP-NewtonD and NIHT only apply for ARMP, it is worth examining their empirical performance on MCP. In our own implementation of these algorithms, we use the set of parameters as suggested by the authors. For SVT, we set the step size $\delta = 1.2\frac{mn}{s}$ and the threshold $\tau = 5\sqrt{mn}$. For SVP-NewtonD, we set the step size $\eta_t = \frac{mn}{1.2s}$. NIHT, IHTSVD and ARNAG-IHT are parameter-free. Finally, we add NAG-IHT with two different fixed step sizes $\beta_k = \frac{1-\sigma}{1+\sigma}$ and $\beta_k = \frac{k-1}{k+2}$ for comparison.

Figure 8.1 illustrates the Frobenius norm of the error matrix as a function of the number of iterations. The dashed lines correspond to the theoretical convergence of IHTSVD (purple) at rate $1-\sigma^2$ and NAG-IHT with step size $\beta_k = \frac{1-\sigma}{1+\sigma}$ (green) at rate $1-\sigma$. As can be seen from the figure, both of the algorithms match the performance predicted in theory. SVT exhibits the slowest convergence due to the conservative nature of proximal-type algorithms. By contrast, all IHT algorithms enjoy a fast convergence at linear rates. Without acceleration, SVP-NewtonD and NIHT are clearly faster than IHTSVD. This can be explained by the fact

that IHTSVD is a special case of SVP when the gradient step size is 1, whereas SVP-NewtonD and NIHT are improved versions of SVP with adaptive step sizes. Notwithstanding, ARNAG-IHT outperforms all other algorithms, asymptotically recovering the convergence rate at $1 - \sigma$. It approaches the "ideal" NAG-IHT with optimal step size in this experiment. Finally, we can observe the periodic behavior of momentum by setting the step size $\beta_k = \frac{k-1}{k+2}$, as experienced in the original version of NAG. However, it can be seen in Fig. 8.1 that this setting does not generally help improve the convergence of plain IHT.

8.6 Conclusion and Future Work

We proposed the use of NAG to boost the performance of IHT for low-rank matrix completion. We analyzed the local convergence of NAG-IHT and established the optimal step size to guarantee faster convergence over plain IHT. We further introduced an adaptive restart algorithm that helps recover the optimal linear rate of convergence in practice. Our numerical evaluation showed evidence that the proposed scheme dramatically improves the performance of IHT for matrix completion problem. Still, understanding when and how our approach works in case the input matrix is noisy and not close to being low-rank is left as an open question for future work.

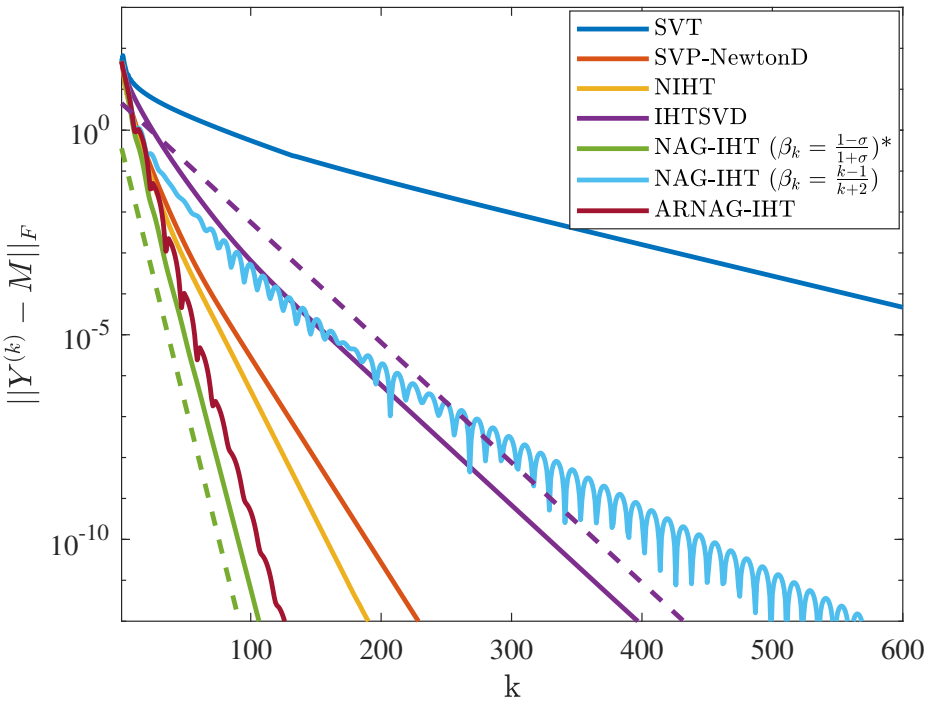


Figure 8.1: The distance to the solution (in log-scale) as a function of the number of iterations for different algorithms (solid) and their corresponding theoretical bounds up to a constant (dashed). The parameters are set to $m = 50, n = 40, r = 3$, and $s = 1000$. Asterisks indicate algorithms using theoretical step sizes that are not available in practice.

8.7 Appendix

8.7.1 Proof of Theorem 8.1

First, we prove the order of singular values is preserved in a neighborhood of the rank-r matrix M. Using Weyl's theorem, we have

$$|\sigma_i(M+\Delta) - \sigma_i| \leq \|\Delta\|_F, \qquad \text{for } 1 \leq i \leq n.$$

For any i such that $\sigma_i > \sigma_{i+1}$: since $\|\Delta\|_F < \frac{\epsilon}{2} \leq \frac{\sigma_i - \sigma_{i+1}}{2}$, the following inequality holds

$$\sigma_{i+1}(M+\Delta) < \sigma_{i+1} + \frac{\sigma_i - \sigma_{i+1}}{2} = \sigma_i - \frac{\sigma_i - \sigma_{i+1}}{2} < \sigma_i(M+\Delta).$$

Thus, the order of singular values is preserved. Moreover, since $\sigma_r(M+\Delta) - \sigma_{r+1}(M+\Delta) > 0$, the top r singular value components are unique and consequently $\mathcal{P}_r(M+\Delta)$ is unique.

Let $M = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ be the rank-r matrix of interest. From matrix perturbation theory [130], we can describe the decomposition of the perturbed matrix

$$M + \Delta = \sum_{i=1}^{r} (\sigma_i + \delta_i)(u_i + \delta u_i)(v_i + \delta v_i)^T + \sum_{i=r+1}^{n} \delta_i (u_i + \delta u_i)(v_i + \delta v_i)^T, \quad (8.5)$$

where $\delta_i, \delta u_i$, and δv_i have norms **in the order of** $O(\|\Delta\|_F)$. Since the top-r singular values of M are preserved under perturbation, we have $\mathcal{P}_r(M+\Delta) =$

$\sum_{i=1}^{r}(\sigma_i+\delta_i)(u_i+\delta u_i)(v_i+\delta v_i)^T$ and (8.5) can be reorganized as

$$\begin{aligned}\mathcal{P}_r(M+\Delta)-M &= \Delta-\sum_{i=r+1}^{n}\delta_i(u_i+\delta u_i)(v_i+\delta v_i)^T \\ &= \Delta-\sum_{i=r+1}^{n}u_i\delta_i v_i^T+O(\|\Delta\|_F^2). \end{aligned} \tag{8.6}$$

Further, substituting $M=\sum_{i=1}^{r}\sigma_i u_i v_i^T$ into (8.5) yields

$$\Delta=\sum_{i=1}^{n}\left(\delta_i u_i v_i^T+\sigma_i\delta u_i v_i^T+\sigma_i u_i\delta {v_i}^T\right)+O(\|\Delta\|_F^2).$$

Then using the orthogonality of u_i, v_i, we can obtain

$$u_i^T\Delta v_i=\delta_i+\sigma_i(u_i^T\delta u_i+\delta {v_i}^T v_i)+O(\|\Delta\|_F^2), \tag{8.7}$$

$$u_i^T\Delta v_j=O(\|\Delta\|_F^2). \tag{8.8}$$

The second term on the RHS can be computed as follows

$$\begin{aligned} & I=\sum_{i=1}^{n}(u_i+\delta u_i)(u_i+\delta u_i)^T \\ \Rightarrow\quad & 1=u_i^Tu_i=1+u_i^T\delta u_i+\delta {u_i}^T u_i+O(\|\Delta\|_F^2) \\ \Rightarrow\quad & u_i^T\delta u_i=O(\|\Delta\|_F^2). \end{aligned}$$

Similarly, we also have $v_i^T\delta v_i=O(\|\Delta\|_F^2)$. Substituting back to (8.7), we get

$\delta_i = u_i^T \Delta v_i + O(\|\Delta\|_F^2)$. Thus, (8.6) can be rewritten as

$$\begin{aligned}\mathcal{P}_r(M+\Delta) - M &= \Delta - \sum_{i=r+1}^{n} u_i u_i^T \Delta v_i v_i^T + O(\|\Delta\|_F^2) \\ &= \Delta - U_2 U_2^T \Delta V_2 V_2^T + O(\|\Delta\|_F^2).\end{aligned}$$

where the last equation stems from (8.8).

8.7.2 Proof of Theorem 8.3

The error matrix can be represented as follows:

$$\begin{aligned}E^{(k)} = Y^{(k)} - M &= \mathcal{P}_{M,\mathcal{S}}\Big(X^{(k)} + \beta(X^{(k)} - X^{(k-1)})\Big) - M \\ &= [(1+\beta)(X^{(k)} - M) - \beta(X^{(k-1)} - M)]_{\mathcal{S}^c} \\ &= (1+\beta)[\mathcal{P}_r(Y^{(k-1)}) - M]_{\mathcal{S}^c} - \beta[\mathcal{P}_r(Y^{(k-2)}) - M]_{\mathcal{S}^c}.\end{aligned}$$

Using a vectorized version of Theorem 8.1, we can reformulate the above equation as

$$e^{(k)} = (1+\beta)(I_d - H)e^{(k-1)} - \beta(I_d - H)e^{(k-2)} + (1+\beta)q(e^{(k-1)}) - \beta q(e^{(k-2)}).$$

where $d = mn - s$, $e^{(k)} = S_c \operatorname{vec}(E^{(k)})$, $H = S_c(V_2 \otimes U_2)(V_2 \otimes U_2)^T S_c^T$ and $q(S_c \operatorname{vec}(\Delta)) = S_c \operatorname{vec}(Q(\Delta))$. By stacking $e^{(k)}$ and $e^{(k-1)}$ together, the recursion

can be rewritten as follows

$$\begin{bmatrix} e^{(k)} \\ e^{(k-1)} \end{bmatrix} = \underbrace{\begin{bmatrix} (1+\beta)(I_d - H) & -\beta(I_d - H) \\ I_d & 0 \end{bmatrix}}_{T} \begin{bmatrix} e^{(k-1)} \\ e^{(k-2)} \end{bmatrix} + \begin{bmatrix} (1+\beta)q(e^{(k-1)}) - \beta q(e^{(k-2)}) \\ 0 \end{bmatrix}.$$

Now, using Lemma 10 in [166], we obtain the upper bound

$$\left\| \begin{bmatrix} e^{(k)} \\ e^{(k-1)} \end{bmatrix} \right\|_2 \leq \big(\rho(T) + o(1)\big)^{k-1} \left\| \begin{bmatrix} e^{(1)} \\ e^{(0)} \end{bmatrix} \right\|_2,$$

where $\rho(T)$ is the spectral radius of T and is equal to the maximum magnitude of any eigenvalue of T.

We compute $\rho(T)$ as follows. Since H is a real symmetric in $\mathbb{R}^{d\times d}$, let $H = U\Lambda U^T$ be the eigenvalue decomposition of H, where U is a unitary matrix and Λ is a diagonal matrix whose entries are the eigenvalues of H:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d = \sigma^2.$$

Define the permutation π as

$$\pi(j) = \begin{cases} 2j - 1 \text{ if } j \leq d, \\ 2j - 2d \text{ otherwise.} \end{cases}$$

Denote the permutation matrix associated with π by P_π. Then, T can be shown to be similar to a block diagonal matrix

$$T \sim P_\pi \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}^T \begin{bmatrix} (1+\beta)(I_d - H) & -\beta(I_d - H) \\ I_d & \mathbf{0} \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} P_\pi^T$$
$$= \begin{bmatrix} T_1 & 0 & \dots & 0 \\ 0 & T_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & T_d \end{bmatrix},$$

where each 2×2 block T_j is of the form

$$\begin{bmatrix} (1+\beta)(1-\lambda_j) & -\beta(1-\lambda_j) \\ 1 & 0 \end{bmatrix}$$

for $j = 1, \dots, mn$. Thus, the eigenvalues of T are also the eigenvalues of all blocks T_j. Finding optimal step size β is equivalent to solving the following problem

$$\min_\beta \max_r |r| \quad \text{such that } r^2 - (1+\beta)(1-\lambda_j)r + \beta(1-\lambda_j) = 0,$$

for some $j \in \{1, \dots, d\}$. Since H is a semi-unitary matrix, we have $\lambda_j \leq 1$ for all j. Each quadratic equation has three cases:

1. If $\Delta = (1-\lambda_j)\Big((1-\lambda_j)(1+\beta)^2 - 4\beta\Big) = 0$ or $\beta = \beta_j^* = \frac{1-\sqrt{\lambda_j}}{1+\sqrt{\lambda_j}}$, then there are two real repeat roots $r_{j1} = r_{j2} = \sqrt{\beta(1-\lambda_j)}$.

2. If $\Delta > 0$ or $\beta < \beta_j^*$, then there are two real distinct roots r_{j1}, r_{j2}. The convergence rate depends on $\max\{|r_{j1}|, |r_{j2}|\}$, which is greater than $\sqrt{|\beta(1-\lambda_j)|}$.

3. If $\Delta < 0$ or $\beta > \beta_j^*$, then there are two conjugate complex roots satisfying $|r_{j2}| = |r_{j2}| = \sqrt{\beta(1-\lambda_j)}$.

In any case, we have $\rho(T) = \max_j |r_j| \geq \sqrt{|\beta(1-\lambda_d)|}$. The equality holds when setting $\beta = \frac{1-\sqrt{\lambda_d}}{1+\sqrt{\lambda_d}}$.

Chapter 9: Local Convergence of the Heavy Ball method in Iterative Hard Thresholding for Low-Rank Matrix Completion[1]

We present a momentum-based accelerated iterative hard thresholding (IHT) for low-rank matrix completion. We analyze the convergence of the proposed Heavy Ball (HB) accelerated IHT near the solution and provide optimal step size parameters that guarantee the fastest rate of convergence. Since the optimal step sizes depend on the unknown structure of the solution matrix, we further propose a heuristic for parameter selection that is inspired by recent results in random matrix theory. Our experiment on a simple matrix completion setting verifies our analysis and illustrates the competitive rate of convergence that can be obtained with the proposed algorithm.

9.1 Introduction

This chapter studies the problem of low-rank matrix completion. Given an $m \times n$ matrix M with low rank r and a set $\mathcal{S} \subset [m] \times [n]$ of its observed entries, where $[m] = \{1, 2, \ldots, m\}$, the goal is to recover the remaining entries of M. Similar to

[1]This work has been published as: Trung Vu and Raviv Raich. "Local Convergence of the Heavy Ball method in Iterative Hard Thresholding for Low-Rank Matrix Completion." In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3417-3421. IEEE, 2019.

sparse recovery, the matrix completion problem (MCP) is shown to be NP-hard [45], considering the non-convexity of the problem rooted in the rank constraint.

In 2009, Candès and Recht [33] achieved a major breakthrough in matrix completion. The authors presented a convex relaxation approach to MCP by replacing the non-convex rank minimization with a (convex) nuclear norm minimization. They showed that one can perfectly recover most low-rank matrices provided the cardinality of $\mathcal{S}$ is sufficiently large. Following this work, a plethora of algorithms have been proposed for low-rank matrix completion via nuclear norm minimization. Among which, first-order methods (e.g., proximal-type algorithms) have grown more attractive due to their simplicity and scalability. However, the conservative nature of the soft thresholding operator associated with such methods often results in slow convergence.

To improve convergence while maintaining scalability, the original non-convex formulation of the problem was revisited. Empirical evidence indicated that iterative approaches to the non-convex rank minimization are faster to converge compared to their convex counterparts. Notwithstanding, theoretical convergence guarantees for such methods are non-trivial and often rely on the Restricted Isometry Property (RIP) of the affine transformations in matrix sensing. Most known examples in this category include iterative hard thresholding (IHT) [104] and alternating minimization (AMMC) [40]. Unfortunately, RIP does not hold for matrix completion even though this problem is a special case of matrix sensing. Thus, recent efforts in understanding algorithms for MCP are limited to probabilistic convergence guarantees [106, 115] or local convergence analysis [46, 120]. More-

over, acceleration techniques have been introduced to improve the performance of IHT in matrix sensing [116, 119]. Under similar assumptions to matrix RIP, the authors provided an analysis of momentum behavior and proved the linear convergence of accelerated IHT. Empirically, the authors of [116] demonstrated a faster convergence of accelerated IHT relative to plain IHT. However, they stated that the sufficient conditions to guarantee such acceleration remain as an open question.

In this work, we develop an accelerated variant of IHT for solving MCP. While the aforementioned approaches to accelerating IHT employ Nesterov's Accelerated Gradient method, we utilize Heavy Ball method due to its faster local convergence. In particular, we provide a theoretical analysis on the local convergence of the proposed algorithm and identify the choice of step sizes that guarantees optimal acceleration. Since it is computationally expensive to perform line search for the momentum parameters, we propose a simple heuristic to approximate the optimal values based on recent results from random matrix theory. Our experiment verifies the convergence rates obtained in our analysis and illustrates the efficiency of the proposed algorithm.

9.2 Notation

Without loss of generality, assume $m \geq n$. Assume the solution matrix $M = U\Sigma V^T$ is a rank-r matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$.

We partition U, Σ, V as follows:

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$$

where $\Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_r)$, $\Sigma_2 = \mathbf{0}$, and U_1, V_1, U_2, V_2 are semi-unitary matrices corresponding to the partition of Σ.

Let $X \in \mathbb{R}^{m \times n}$ be an arbitrary matrix. We define the rank-r projection $\mathcal{P}_r$ as $\mathcal{P}_r(X) = \sum_{i=1}^{r} \sigma_i(X) u_i(X) v_i(X)^T$, where $\sigma_i(X)$, $u_i(X)$, and $v_i(X)$ are the i-th singular value, column vector, and row vector, of X, respectively. This projection produces the best rank-r approximation of X [61] and it is unique if either $\sigma_r(X) > \sigma_{r+1}(X)$ or $\sigma_r(X) = 0$. Further, we denote the cardinality of $\mathcal{S}$ by s. The sampling operator $X_{\mathcal{S}}$ is given by

$$[X_{\mathcal{S}}]_{ij} = \begin{cases} X_{ij} & \text{if } (i,j) \in \mathcal{S}, \\ 0 & \text{if } (i,j) \in \mathcal{S}^c. \end{cases}$$

where $\mathcal{S}^c$ is the complement of $\mathcal{S}$. Let $\hat{\mathcal{S}}^c = \{i + m(j-1) \mid (i,j) \in \mathcal{S}^c\}$. We define $S_c \in \mathbb{R}^{(mn-s) \times mn}$ as the row selection matrix obtained by selecting a subset of rows corresponding to the elements of $\hat{\mathcal{S}}^c$ from the $mn \times mn$ identity matrix.

Algorithm 9.1 Iterative Hard Thresholding

1: $X^{(0)} = M_{\mathcal{S}}$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad X^{(k)} = \mathcal{P}_r\big(X^{(k-1)} - \alpha_k[X^{(k-1)} - M]_{\mathcal{S}}\big)$

9.3 Background

Iterative hard thresholding for matrix recovery was first introduced by Jain et. al. [104] and quickly became a very attractive method for solving this problem, thanks to its simplicity and efficiency over the proximal-type algorithms [27]. Despite the successful development in theoretical analyses of IHT for *matrix sensing* [116,201], there has been little progress in understanding the convergence of IHT for *low-rank matrix completion*. The lack of RIP guarantees for MCP leaves the global convergence of IHT for MCP as an open question. Nonetheless, empirical performance analysis of the algorithm often shows linear convergence of the approach. Hence, there have been efforts to establish local convergence guarantees [46,120]. Notably, the authors of [46] showed that the local rate of convergence of MCP-IHT can be described in a closed-form. We review the IHT algorithm for matrix completion in Algorithm 9.1 and restate the local convergence results in Theorem 9.1 and Theorem 9.2, using our aforementioned notations for consistency.

Theorem 9.1. *(Rephrased from [46]) Let $\Delta \in \mathbb{R}^{m\times n}$ be a perturbation matrix such that $\|\Delta\|_F < \frac{\epsilon}{2}$, where $\epsilon = \min_{\sigma_i > \sigma_{i+1}} \{\sigma_i - \sigma_{i+1}\}$. Then the rank-r projection of $M + \Delta$ is given by*

$$\mathcal{P}_r(M + \Delta) = M + \Delta - U_2 U_2^T \Delta V_2 V_2^T + Q(\Delta) \tag{9.1}$$

where $Q : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ *satisfies* $\|Q(\Delta)\|_F = O(\|\Delta\|_F^2)$.

Note that (9.1) can be vectorized as $\operatorname{vec}\big(\mathcal{P}_r(M+\Delta)\big) = \operatorname{vec}(M) + \big(I_{mn} - (V_2 \otimes U_2)(V_2 \otimes U_2)^T\big)\operatorname{vec}(\Delta) + q(\operatorname{vec}(\Delta))$, where $q(\operatorname{vec}(\Delta)) = \operatorname{vec}(Q(\Delta))$. Denote the error vector $e^{(k)} = S_c \operatorname{vec}(X^{(k)} - M)$. Then considering Algorithm 9.1 with a unit step size ($\alpha_k = 1$), one can show a recursion of the error vector as follows

$$e^{(k)} = (I_{mn-s} - H)e^{(k-1)} + q(e^{(k-1)})$$

where $H = S_c(V_2 \otimes U_2)(V_2 \otimes U_2)^T S_c^T$. Further, let $L = \lambda_{\max}(H)$ and $\mu = \lambda_{\min}(H)$ be the largest and smallest eigenvalues of H, respectively. Since H is positive semi-definite and S_c, V_2, U_2 are semi-unitary matrices, it holds that $0 \le \mu \le L \le 1$.

Theorem 9.2. *(Rephrased from [46]) If* $\mu > 0$*, then Algorithm 9.1 with a unit step size converges to* M *locally at a linear rate* $1-\mu$*. In other words, there exists a neighborhood* $\mathcal{E}(M)$ *of* M *and a constant* C *such that if* $X^{(0)} \in \mathcal{E}(M)$*, then*

$$\big\|X^{(k)} - M\big\|_F \le C\big(1-\mu\big)^k \big\|X^{(0)} - M\big\|_F .$$

Interestingly, the convergence rate $1-\mu$ depends only on the solution M and the set of observed entries $\mathcal{S}$. It is also noteworthy that similar local linear convergence has been studied later in [120]. However, there is no explicit formulation of the convergence rate specified by the authors.

To gain intuition into accelerated IHT, let us start with classic results on the convergence of first-order methods for minimizing *convex quadratic functions*. In

Table 9.1: Parameter selection and convergence rate of different first-order methods for minimizing a convex quadratic function $f(x) = \frac{1}{2}x^T Ax + b^T x + c$, where $x \in \mathbb{R}^d$ and $\mu I_d \preceq A \preceq L I_d$. Asterisks indicate algorithms with optimal fixed step sizes. The last column describes the proportional numbers of iterations needed to reach a relative accuracy ϵ, i.e., $\left\|x^{(k)} - x^*\right\|_2 \leq \epsilon \left\|x^{(0)} - x^*\right\|_2$. All algorithms share the same computational complexity per iteration.

Method	**Update at each iteration**	**Step size selection**	**Rate**	**#Iters. needed**
Gradient	$x^{(k)} = x^{(k-1)} - \alpha \nabla f(x^{(k-1)})$	$\alpha = \frac{1}{L}$	$1 - \frac{\mu}{L}$	$\frac{L}{\mu}\log(1/\epsilon)$
Gradient*		$\alpha = \frac{2}{L+\mu}$	$1 - \frac{2\mu}{L+\mu}$	$\frac{1}{2}(\frac{L}{\mu}+1)\log(1/\epsilon)$
Nesterov	$y^{(k)} = x^{(k-1)} - \alpha \nabla f(x^{(k-1)})$ $x^{(k)} = y^{(k-1)} + \beta(y^{(k-1)} - y^{(k-2)})$	$\alpha = \frac{1}{L}, \beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$	$1 - \frac{\sqrt{\mu}}{\sqrt{L}}$	$\sqrt{\frac{L}{\mu}}\log(1/\epsilon)$
Nesterov*		$\alpha = \frac{4}{3L+\mu}, \beta = \frac{\sqrt{3L+\mu}-2\sqrt{\mu}}{\sqrt{3L+\mu}+2\sqrt{\mu}}$	$1 - 2\frac{\sqrt{\mu}}{\sqrt{3L+\mu}}$	$\frac{1}{2}\sqrt{3\frac{L}{\mu}+1}\log(1/\epsilon)$
Heavy Ball*	$x^{(k)} = x^{(k-1)} - \alpha \nabla f(x^{(k-1)})$ $+ \beta(x^{(k-1)} - x^{(k-2)})$	$\alpha = \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2, \beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$	$1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$	$\frac{1}{2}(\sqrt{\frac{L}{\mu}}+1)\log(1/\epsilon)$

Table 9.1, the parameter selection is optimal in the sense that no other choice of fixed step sizes achieves faster convergence rate (see details in [128]). We list methods in ascending order of the convergence rate. In fact, Heavy Ball method not only has the fastest rate but also achieves the lower bound on convergence rate for any first-order methods for minimizing μ-strongly convex, L-smooth functions [158]. Extending these results to study the local convergence of those algorithms for optimizing a non-convex function, one could argue that the objective function can be well approximated by a quadratic inside the region near the optimum. Hence, we consider an HB-variant of MCP-IHT and analyze its local convergence behavior.

9.4 Main results

We begin this section by a brief discussion on parameter selection for plain IHT. In [104], the authors suggested an empirical choice of $\alpha_k = \frac{mn}{(1+\delta)s}$, where δ is a constant determined from experiments. To further investigate the step-size selection, we examine the local convergence rate for Algorithm 9.1 and obtain the optimal step size in the following theorem.

Theorem 9.3. *If $\mu > 0$, then Algorithm 9.1 with step size $\alpha_k = \frac{2}{L+\mu}$ converges to M locally at a linear rate $1 - \frac{2\mu}{L+\mu}$. In other words, there exists a neighborhood $\mathcal{E}(M)$ of M and a constant C such that if $X^{(0)} \in \mathcal{E}(M)$, then*

$$\left\|X^{(k)} - M\right\|_F \leq C\big(1 - \frac{2\mu}{L+\mu}\big)^k \left\|X^{(0)} - M\right\|_F .$$

Algorithm 9.2 HB-IHT

1: $X^{(0)} = X^{(1)} = M_{\mathcal{S}}$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad X^{(k+1)} = \mathcal{P}_r\big(X^{(k)} - \alpha_k[X^{(k)} - M]_{\mathcal{S}}\big) + \beta_k(X^{(k)} - X^{(k-1)})$

Although the optimal step size in Theorem 9.3 is similar to the classical result in Table 9.1, we note that the analysis addresses the issue on the non-convex nature of the rank-r projection.

9.4.1 HB-IHT

Similar to the classic Heavy Ball method, we propose an accelerated algorithm that adds a momentum term to the update in plain IHT (see Algorithm 9.2). This simple modification to plain IHT maintains the computational complexity of the algorithm with one additional step of calculating the difference matrix. On the other hand, the local rate of convergence can be improved significantly. Theorem 9.4 characterizes the local convergence of HB-IHT by providing the optimal parameter selection that guarantees improvement over plain IHT.[2]

Theorem 9.4. *If $\mu > 0$, then Algorithm 9.2 with step sizes $\alpha_k = \big(\frac{2}{\sqrt{L}+\sqrt{\mu}}\big)^2, \beta_k = \big(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\big)^2$ converges to M locally at a linear rate $1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. In other words, there exists a neighborhood $\mathcal{E}(M)$ of M and a constant C such that if $X^{(0)} \in \mathcal{E}(M)$, then*

$$\big\|X^{(k)} - M\big\|_F \leq C\big(1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\big)^k \big\|X^{(0)} - M\big\|_F .$$

[2]The proofs of Theorem 9.1, 9.3 and 9.4 are given in the Appendix at the end of this chapter.

Further, this is the optimal rate among all fixed α, β.

It is noteworthy that despite the operation of non-convex rank-r projections, we still end up with the similar result given in Table 9.1, thanks to the approximation of $\mathcal{P}_r$ given in (9.1).

9.4.2 A Practical Guide to Parameter Selection

Step size selection is critical to the performance of HB-IHT in practice. In this section, we propose a simple heuristic to determine the values of α_k and β_k in Algorithm 9.2 with no prior knowledge about L and μ. The idea is to exploit the special structure of H in order to estimate its extreme eigenvalues. We express this matrix in form of $H = WW^T$, where

$$W = S_c(V_2 \otimes U_2) = S_c(V \otimes U)(S_{2V} \otimes S_{2U})^T,$$

and $S_{2U} \in \mathbb{R}^{(m-r)\times m}$, $S_{2V} \in \mathbb{R}^{(n-r)\times n}$ are row selection matrices. Note that W is a submatrix of the Kronecker product $V \otimes U$ with the row ratio $p = 1 - \frac{s}{mn}$ and the column ratio $q = (1 - \frac{r}{m})(1 - \frac{r}{n})$. In this representation, the structure of H is closely related to the MANOVA random matrix ensemble, and more interestingly, the limiting density of its eigenvalues is identified by Watcher [219], dating back to the early 1980s. In his study, Watcher showed that as the size of a MANOVA matrix with parameters (p, q) approaches infinity, its empirical spectral distribution (ESD) converges to a deterministic probability measure supported on the interval

$[\lambda^-, \lambda^+] \cup \{0, 1\}$, where $\lambda^{\pm} = \left(\sqrt{p(1-q)} \pm \sqrt{q(1-p)}\right)^2$. Recently, similar result was also found by Raich and Kim [171] for the truncation of random unitary matrices. Moreover, Farrell and Nadakuditi [63] extended the results from Haar (uniformly) distributed unitary matrices to Kronecker product case. The authors proved that the ESD of random matrices of the form $\Pi_1(U \otimes V)\Pi_2(U \otimes V)^* \Pi_1$, where Π_1, Π_2 are orthogonal projections of ranks pn and qn, respectively, also converges to the same limiting distribution. Considering H to be an instance of this case, we conjecture that its spectral distribution will be close to the aforementioned. In particular,

1. if $p < q$, then H has no zero eigenvalue and the smallest eigenvalue of H is close to λ^- with high probability,

2. if $p + q > 1$, then H has unit eigenvalue and the largest eigenvalue of H is 1.

It is worthwhile to note that both conditions usually hold in practice when q is rather close to 1. Hence, we propose the following estimation of L and μ:

$$\hat{L} = 1, \quad \hat{\mu} = \left(\sqrt{q(1-p)} - \sqrt{p(1-q)}\right)^2. \tag{9.2}$$

Empirically, we observe this heuristic significantly outperforms plain IHT in terms of convergence. However, understanding when and how it works would involve the non-asymptotic theory of random matrices [180]. For instance, characterizing the variance of extreme eigenvalues, i.e., difference between μ and $\hat{\mu}$, in case of Kronecker unitary matrices is much more challenging than their Haar-distributed

counterparts. Our experiments suggest that they tend to have wider fluctuations. We leave this analysis for future direction.

9.5 Numerical Evaluation

This section presents an empirical evaluation of several methods for low-rank matrix completion including the proposed approach. First, we generate a low-rank solution matrix $M \in \mathbb{R}^{m\times n}$ by taking the product of an $m \times r$ matrix and an $r \times n$ matrix, each having i.i.d. normally distributed entries. Next, we sample the observation set $\mathcal{S}$ uniformly at random. In our experiment, we choose $m = 50$, $n = 40$, $r = 3$, and $s = 1000$. For comparison, we consider the following methods: SVT [27], SVP [104], IHTSVD [46] and AMMC [106]. Although the convergence guarantee of SVP does not hold for MCP in general, it is interesting to compare its empirical performance with optimal step size given in Theorem 9.3. In our own implementation of these algorithms, we use the set of parameters as suggested by the authors. For the proximal-type SVT algorithm, we set the step size $\delta = 1.2\frac{mn}{s}$ and the threshold $\tau = 5\sqrt{mn}$. For SVP, we set the step size $\eta_t = \frac{mn}{1.2s}$. IHTSVD and AMMC are parameter-free. Finally, we add HB-IHT with the aforementioned theoretical optimal step sizes and heuristic step sizes for comparison.

Figure 9.1 illustrates the Frobenius norm of the error matrix as a function of the number of iterations. The dashed lines correspond to the theoretical convergence of IHTSVD (purple) at rate $1 - \mu$, optimal step size SVP (yellow) at rate $1 - \frac{2\mu}{L+\mu}$ and optimal step size HB-IHT (green) at rate $1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. These three

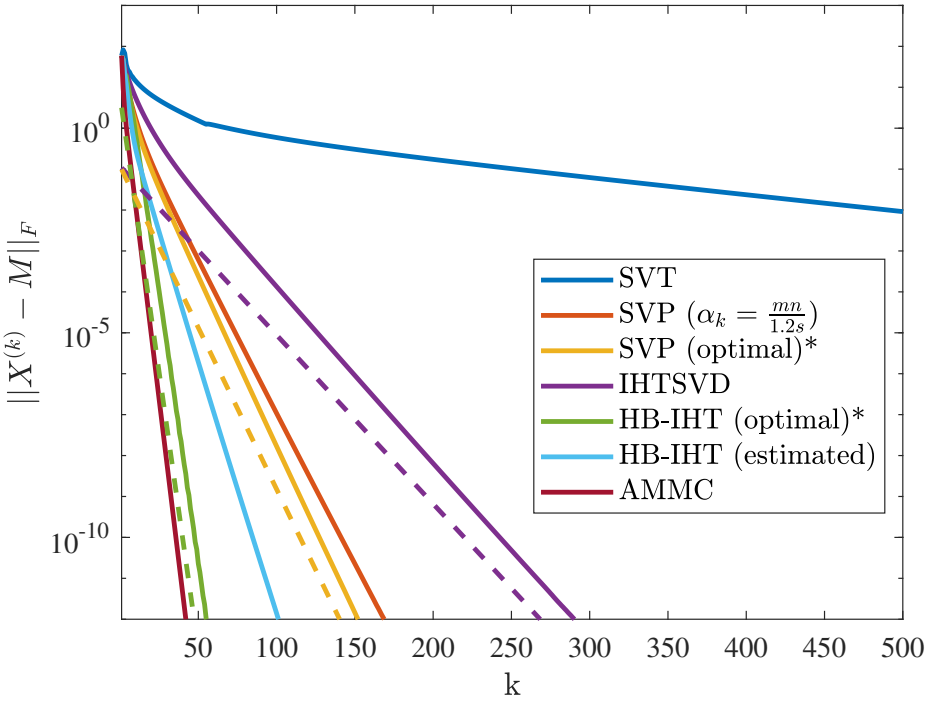


Figure 9.1: The distance to the solution (in log-scale) as a function of the iteration number for various algorithms (solid) and their corresponding theoretical bounds up to a constant (dashed). Asterisks indicate algorithms using theoretical step sizes that are not available in practice. All algorithms share the same computational complexity per iteration except AMMC.

algorithms certainly match the performance predicted in theory. SVT exhibits the slowest convergence as expected from our foregoing discussion. By contrast, all IHT algorithms enjoy the linear convergence. Without acceleration, SVP with step size either $\frac{mn}{1.2s}$ or $\frac{2}{L+\mu}$ is clearly faster than IHTSVD. Nevertheless, HB-IHT with estimated step sizes outperforms all plain IHT algorithms, yet still slower than HB-IHT with theoretically-optimal step sizes. Finally, we compare the performance of HB-IHT with optimal step sizes with AMMC, which is shown to converge linearly at rate faster than 1/4 in [106]. While our accelerated algorithm obtains a comparable rate, it requires significantly less computation per iteration thanks to the recent breakthroughs in k-SVD algorithms [5], i.e., the iteration complexity for HB-IHT is $O(mnr + poly(1/\epsilon))$, compared to $O(sm^2r^2 + m^3r^3)$ for AMMC as claimed in [106].

9.6 Conclusion and Future Work

To summarize, we introduced the use of Heavy Ball method to significantly accelerate IHT for low-rank matrix completion. We analyzed the local convergence of HB-IHT and established the optimal step sizes to guarantee better performance over plain IHT. We further provided evidence that these optimal values can be approximated by a simple calculation in practice. Our experiment verified the analysis and demonstrated the efficiency of the proposed algorithm. Study of our approach in the noisy case is left for an extended version of this work.

9.7 Appendix

9.7.1 Proof of Theorem 9.3

Vectorizing Theorem 9.1 yields

$$\operatorname{vec}(\mathcal{P}_r(M+\Delta)-M) = (I_{mn} - (V_2 \otimes U_2)(V_2 \otimes U_2)^T)\operatorname{vec}(\Delta) + q(\operatorname{vec}(\Delta)) \quad (9.3)$$

where $q(\operatorname{vec}(\Delta)) = \operatorname{vec}(Q(\Delta))$. From the IHT update, the error matrix is

$$\begin{aligned} E^{(k)} &= X^{(k)} - M \\ &= \mathcal{P}_r\Big(X^{(k-1)} - \alpha[X^{(k-1)} - M]_{\mathcal{S}}\Big) - M \\ &= \mathcal{P}_r\Big(M + E^{(k-1)} - \alpha[E^{(k-1)}]_{\mathcal{S}}\Big) - M. \end{aligned}$$

From (9.3), we have

$$\begin{aligned} e^{(k)} &= \operatorname{vec}(E^{(k)}) \\ &= (I_{mn} - WW^T)\operatorname{vec}(E^{(k-1)} - \alpha[E^{(k-1)}]_{\mathcal{S}}) + q\big(\operatorname{vec}(E^{(k-1)} - \alpha[E^{(k-1)}]_{\mathcal{S}})\big) \\ &= (I_{mn} - WW^T)(I_{mn} - \alpha S^T S)e^{(k-1)} + q\big((I_{mn} - \alpha S^T S)e^{(k-1)}\big) \\ &= (I_{mn} - WW^T)((1-\alpha)I_{mn} + \alpha S_c^T S_c)e^{(k-1)} + C_1 q(e^{(k-1)}) \\ &= \left(I_{mn} - \alpha\Big((I_{mn} - WW^T)(I_{mn} - S_c^T S_c) + \frac{1}{\alpha}WW^T\Big)\right)e^{(k-1)} + C_1 q(e^{(k-1)}), \end{aligned}$$

where $W = V_2 \otimes U_2 \in \mathbb{R}^{mn\times(m-r)(n-r)}$ and C_1 is some positive constant. Now, denote

$$Z_\alpha = (I_{mn} - WW^T)(I_{mn} - S_c^T S_c) + \frac{1}{\alpha} WW^T.$$

Using Lemma 10 in [166], we obtain the upper bound

$$\left\|e^{(k)}\right\|_2 \leq \left(\rho(I_{mn} - \alpha Z_\alpha) + o(1)\right)^k \left\|e^{(0)}\right\|_2,$$

where $\rho(I_{mn}-\alpha Z_\alpha)$ is the spectral radius of $I_{mn}-\alpha Z_\alpha$ and is equal to the maximum magnitude of any eigenvalue of $I_{mn} - \alpha Z_\alpha$. Denote $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{mn} \geq 0$ are eigenvalues of Z_α and assume that Z_α is diagonalizable. Then, finding optimal step size α is equivalent to solving the following problem

$$\min_\alpha \max_{1\leq j\leq mn} |1 - \alpha\lambda_j|. \tag{9.4}$$

The solution of the optimization problem (9.4) is given by $\alpha = \frac{2}{\lambda_1+\lambda_{mn}}$ and the optimal rate $\rho(I - \alpha Z_\alpha) = \frac{\lambda_1-\lambda_{mn}}{\lambda_1+\lambda_{mn}}$. Now, using the following lemma to simplify the calculation of λ_j, we obtain $\lambda_1 = L$ and $\lambda_{mn} = \mu$.

Lemma 9.1. *For any $\lambda \in \Lambda(Z_\alpha)$, we have either $\lambda = \frac{1}{\alpha}$ or $\lambda = 1$ or $\lambda \in \Lambda(H)$, where $H = S_c WW^T S_c^T$.*

Proof. For any $\lambda \in \Lambda\Big((I_{mn} - WW^T)(I_{mn} - S_c^T S_c) + \frac{1}{\alpha} WW^T\Big)$, there exists $v \in$

$\mathcal{C}^{mn}, v \neq \mathbf{0}$ such that

$$\Big((I_{mn} - WW^T)(I_{mn} - S_c^T S_c) + \frac{1}{\alpha} WW^T\Big)v = \lambda v. \tag{9.5}$$

Left-multiplying both sides with $(I_{mn} - WW^T)$ and recall that $W^T W = I_{(m-r)(n-r)}$, we have

$$(I_{mn} - WW^T)(I_{mn} - S_c^T S_c)v = \lambda(I_{mn} - WW^T)v.$$

Substituting back into (9.5), we get

$$\lambda(I_{mn} - WW^T)v + \frac{1}{\alpha} WW^T v = \lambda v \qquad \Rightarrow \qquad (\frac{1}{\alpha} - \lambda)WW^T v = \mathbf{0}.$$

Hence, we have either $\lambda = \frac{1}{\alpha}$ or $WW^T v = \mathbf{0}$. In the later case, we can substitute into (9.5) again to obtain

$$\lambda v = (I_{mn} - WW^T)(I_{mn} - S_c^T S_c)v = (I_{mn} - S_c^T S_c + WW^T S_c^T S_c)v. \tag{9.6}$$

Left-multiplying both sides with S_c and recall that $S_c S_c^T = I_{mn-s}$, we have

$$S_c WW^T S_c^T (S_c v) = \lambda(S_c v).$$

If $S_c v = \mathbf{0}$, then plugging into (9.6) yields $\lambda = 1$. Otherwise, we have $\lambda \in \Lambda(S_c WW^T S_c^T)$. This completes our proof of the lemma. □

9.7.2 Proof of Theorem 9.4

The error matrix can be represented as follows

$$E^{(k+1)} = X^{(k+1)} - M = \mathcal{P}_r\Big(X^{(k)} - \alpha[X^{(k)} - M]_{\mathcal{S}}\Big) + \beta(X^{(k)} - X^{(k-1)}) - M$$
$$= \Big(\mathcal{P}_r\big(M + E^{(k)} - \alpha[E^{(k)}]_{\mathcal{S}}\big) - M\Big) + \beta\Big(E^{(k)} - (E^{(k-1)}\Big).$$

Similarly to Theorem 10.1, we can vectorize the above equation as

$$e^{(k+1)} = \Big(I_{mn} - \alpha Z_{\alpha}\Big)e^{(k)} + \beta(e^{(k)} - e^{(k-1)}) + C_1 q(e^{(k)}).$$

By stacking $e^{(k+1)}$ and $e^{(k)}$ together, the recursion can be rewritten as follows

$$\begin{bmatrix} e^{(k+1)} \\ e^{(k)} \end{bmatrix} = \underbrace{\begin{bmatrix} (1+\beta)I_{mn} - \alpha Z_{\alpha} & -\beta I_{mn} \\ I_{mn} & \mathbf{0} \end{bmatrix}}_{T} \begin{bmatrix} e^{(k)} \\ e^{(k-1)} \end{bmatrix} + \begin{bmatrix} C_1 q(e^{(k)}) \\ 0 \end{bmatrix}.$$

Now, using Lemma 10 in [166], we obtain the upper bound

$$\left\| \begin{bmatrix} e^{(k+1)} \\ e^{(k)} \end{bmatrix} \right\|_2 \leq \big(\rho(T) + o(1)\big)^k \left\| \begin{bmatrix} e^{(1)} \\ e^{(0)} \end{bmatrix} \right\|_2$$

where $\rho(T)$ is the spectral radius of T. Assume that Z_α is diagonalizable, then T is similar to a block diagonal matrix with 2×2 block T_j of the form [3]

$$\begin{bmatrix} 1+\beta-\alpha\lambda_j & -\beta \\ 1 & 0 \end{bmatrix}$$

for $j = 1, \ldots, mn$. Thus, the eigenvalues of T are also the eigenvalues of all blocks T_j. Finding optimal step size β is equivalent to solving the following problem

$$\min_{\alpha,\beta} \max |r| \text{ such that } r^2 - (1+\beta-\alpha\lambda_j)r + \beta = 0, \text{ for some } j \in \{1, \ldots, mn\}. \tag{9.7}$$

Since $\Delta = (1+\beta-\alpha\lambda)^2 - 4\beta$, it is easy to verify that

- if $\Delta \leq 0$, then $|\sigma_1| = |\sigma_2| = \sqrt{|\beta|}$,
- if $\Delta > 0$, then $\max\{|\sigma_1|, |\sigma_2|\} > \sqrt{|\beta|}$.

The optimization (9.7) becomes

$$\begin{aligned} & \min_{\alpha,\beta} \sqrt{\beta} \qquad \text{s.t. } (1+\beta-\alpha\lambda_j)^2 - 4\beta \leq 0 \text{ for all } 1 \leq j \leq mn \\ \Leftrightarrow \quad & \min_{\alpha,\beta} \max_j \left|1 - \sqrt{\alpha\lambda_j}\right| \qquad \text{s.t. } \beta \geq (1-\sqrt{\alpha\lambda_j})^2 \text{ for all } 1 \leq j \leq mn. \end{aligned} \tag{9.8}$$

[3] This is shown by performing a change of basis on orthogonal space of H, following by permutations on rows and columns.

The solution of the optimization problem (9.8) is given by

$$\alpha = \Big(\frac{2}{\sqrt{\lambda_1} + \sqrt{\lambda_{mn}}}\Big)^2, \qquad \beta = \Big(\frac{\sqrt{\lambda_1} - \sqrt{\lambda_{mn}}}{\sqrt{\lambda_1} + \sqrt{\lambda_{mn}}}\Big)^2.$$

Finally, we obtain the optimal rate $\rho(T) = \frac{\sqrt{\lambda_1}-\sqrt{\lambda_{mn}}}{\sqrt{\lambda_1}+\sqrt{\lambda_{mn}}}$. Now, using Lemma 9.1, we obtain $\lambda_1 = L$ and $\lambda_{mn} = \mu$.

Chapter 10: Exact Linear Convergence Rate Analysis for Low-Rank Symmetric Matrix Completion via Gradient Descent[1]

Factorization-based gradient descent is a scalable and efficient algorithm for solving low-rank matrix completion. Recent progress in structured non-convex optimization has offered global convergence guarantees for gradient descent under certain statistical assumptions on the low-rank matrix and the sampling set. However, while the theory suggests gradient descent enjoys fast linear convergence to a global solution of the problem, the universal nature of the bounding technique prevents it from obtaining an accurate estimate of the rate of convergence. This chapter performs a local analysis of the exact linear convergence rate of gradient descent for factorization-based symmetric matrix completion. Without any additional assumptions on the underlying model, we identify the deterministic condition for local convergence guarantee for gradient descent, which depends only on the solution matrix and the sampling set. More crucially, our analysis provides a closed-form expression of the asymptotic rate of convergence that matches exactly with the linear convergence observed in practice. To the best of our knowledge,

[1]This work has been published as: Trung Vu and Raviv Raich. "Exact Linear Convergence Rate Analysis for Low-Rank Symmetric Matrix Completion via Gradient Descent." In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3240-3244. IEEE, 2021.

our result is the first one that offers the exact linear convergence rate of gradient descent for matrix factorization in Euclidean space for matrix completion.

10.1 Introduction

Matrix completion is the problem of recovering a low-rank matrix from a sampling of its entries. In machine learning and signal processing, it has a wide range of applications including collaborative filtering [176], system identification [137] and dimension reduction [31]. In the era of big data, matrix completion has been proven to be an efficient and powerful framework to handle the enormous amount of information by exploiting low-rank structure of the data matrix.

Let $\boldsymbol{M} \in \mathbb{R}^{n\times m}$ be a rank r matrix with $1 \leq r \leq \min(n, m)$, and $\Omega = \{(i, j) \mid M_{ij}$ is observed$\}$ be an index subset of cardinality s such that $s \leq nm$. The goal is to recover the unknown entries of $\boldsymbol{M}$. Matrix completion can be formulated as a linearly constrained rank minimization or a rank-constrained least squares problem [33]. Two popular approaches to solving matrix completion are convex relaxation via nuclear norm and non-convex factorization. The former approach, motivated by the success of compressed sensing, replaces the matrix rank by its convex surrogate (the nuclear norm). Extensive work on designing convex optimization algorithms with guarantees can be found in [27,33,107,145,205]. While on the theoretical side, the solutions of the relaxed problem and the original problem can be shown to coincide with high probability, on the practical side, computational complexity concerns diminish the applicability of these algorithms. When the size

of the matrix grows rapidly, storing and optimizing over a matrix variable become computationally expensive and even infeasible. In addition, it is evident this approach suffers from slow convergence [117, 213]. In the second approach, the original rank-constrained optimization is studied. Interestingly, by reparametrizing the $n \times m$ matrix as the product of two smaller matrices $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Y}^\top$, for $\boldsymbol{X} \in \mathbb{R}^{n \times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{m \times r}$, the resulting equivalent problem is unconstrained and more computationally efficient to solve [26]. Although this problem is non-convex, recent progress shows that for such problem any local minimum is also a global minimum [76, 199]. Thus, basic optimization algorithms such as gradient descent [43, 144, 199] and alternating minimization [40, 91, 92, 106] can provably solve matrix completion under a specific sampling regime. Alternatively, the original rank-constrained optimization problem can be solved without the aforementioned reparameterization via the truncated singular value decomposition [46, 79, 104, 105, 201, 213, 214].

Among the aforementioned algorithms, let us draw our attention to the gradient descent method due to its outstanding simplicity and scalability. The first global convergence guarantee is attributed to Sun and Luo [199]. The authors proved that gradient descent with appropriate regularization can converge to the global optima of a factorization-based formulation at a linear rate. Later on, Ma *et. al.* [144] proposed that even in the absence of explicit regularization, gradient descent recovers the underlying low-rank matrix by implicitly regularizing its iterates. The aforementioned results, while establishing powerful guarantees on the convergence behavior of gradient descent, impose several limitations. For some methods, the linear convergence rate depends on constants that are not in closed-

form and are hard to verify in numerical experiments even when the underlying matrix is known. Second, a solution-independent analysis of the convergence rate typically offers a loose bound when considered for a specific solution. Third, the global convergence analysis requires certain assumptions on the underlying model which largely restrict the setting of the matrix completion problem in practice. Among such assumptions, one would consider the incoherence of the target matrix, the randomness of the sampling set, and the fact that the rank r and the condition number of $\boldsymbol{M}$ are small constants as $n, m \to \infty$.

To address these issues, we consider the local convergence analysis of gradient descent for factorization-based matrix completion. In the scope of this chapter, we restrict our attention to the symmetric case. We identify the condition for linear convergence of gradient descent that depends only on the solution $\boldsymbol{M}$ and the sampling set Ω. In addition, we provide a closed-form expression for the asymptotic convergence rate that matches well with the convergence of the algorithm in practice. The proposed analysis does not require an asymptotic setting for matrix completion, e.g., large matrices of small rank. We believe that our analysis can be useful in both theoretical and practical aspects of the matrix completion problem.

10.2 Gradient Descent for Matrix Completion

Notations Throughout the chapter, we use the notations $\|\cdot\|_F$ and $\|\cdot\|_2$ to denote the Frobenius norm and the spectral norm of a matrix, respectively. On the other hand, $\|\cdot\|_2$ is used on a vector to denote the Euclidean norm. Boldfaced

symbols are reserved for vectors and matrices. In addition, the $t \times t$ identity matrix is denoted by $\boldsymbol{I}_t$. $\otimes$ denotes the Kronecker product between two matrices, and vec$(\cdot)$ denotes the vectorization of a matrix by stacking its columns on top of one another. Let $\boldsymbol{X}$ be some matrix and $\boldsymbol{F}(\boldsymbol{X})$ be a matrix-valued function of $\boldsymbol{X}$. Then, for some positive number k, we use $\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{\mathcal{O}}(\|\boldsymbol{X}\|_F^k)$ to imply $\lim_{\delta \to 0} \sup_{\|\boldsymbol{X}\|_F = \delta} \|\boldsymbol{F}(\boldsymbol{X})\|_F / \|\boldsymbol{X}\|_F^k < \infty$.

We begin by introducing the low-rank matrix completion problem. For simplicity, we focus on the *symmetric* case where $\boldsymbol{M}$ is an $n \times n$ positive semi-definite (PSD) matrix of rank r and the sampling set Ω is symmetric.[2] Assume the rank-r economy version of the eigendecomposition of $\boldsymbol{M}$ is given by

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top,$$

where $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ is a semi-orthogonal matrix and $\boldsymbol{\Lambda} \in \mathbb{R}^{r \times r}$ is a diagonal matrix containing r non-zero eigenvalues of $\boldsymbol{M}$, i.e., $\lambda_1 \geq \ldots \geq \lambda_r > 0$. Since $\boldsymbol{M}$ can be represented as

$$\boldsymbol{M} = (\boldsymbol{U}\boldsymbol{\Lambda}^{1/2})(\boldsymbol{U}\boldsymbol{\Lambda}^{1/2})^\top,$$

we can write $\boldsymbol{M} = \boldsymbol{X}^* \boldsymbol{X}^{*\top}$, such that $\boldsymbol{X}^* = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2} \in \mathbb{R}^{n \times r}$. Therefore, the factorization-based formulation for matrix completion can be described using the

[2] If the sampling set is not symmetric, one can symmetrize it by adding (j, i), for any $(i, j) \in \Omega$, to Ω since $M_{ji} = M_{ij}$.

Algorithm 10.1 (Non-convex) Gradient Descent

Require: $\boldsymbol{X}^0$, $\mathcal{P}_\Omega(\boldsymbol{M})$, η
Ensure: $\{\boldsymbol{X}^k\}$
1: **for** $k = 0, 1, 2, \ldots$ **do**
2: $\quad \boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta \mathcal{P}_\Omega\big(\boldsymbol{X}^k \boldsymbol{X}^{k\top} - \boldsymbol{M}\big)\boldsymbol{X}^k$

following non-convex optimization:

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \frac{1}{4} \sum_{(i,j) \in \Omega} \big([\boldsymbol{X}\boldsymbol{X}^\top]_{ij} - M_{ij}\big)^2. \tag{10.1}$$

Denote $\mathcal{P}_\Omega : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ the projection onto the set of matrices supported in Ω, i.e.,

$$[\mathcal{P}_\Omega(\boldsymbol{Z})]_{ij} = \begin{cases} Z_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

The objective function in (10.1) is rewritten as $f(\boldsymbol{X}) = \frac{1}{4} \big\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M})\big\|_F^2$. The gradient of $f(\boldsymbol{X})$ is given by

$$\nabla f(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M})\boldsymbol{X}. \tag{10.2}$$

Starting from an initial $\boldsymbol{X}^0$ (usually through spectral initialization [144]), the gradient descent algorithm (see Algorithm 10.1) simply updates the value of $\boldsymbol{X}$ by taking steps proportional to the negative of the gradient $\nabla f(\boldsymbol{X})$.

10.3 Local Convergence Analysis

This section presents the local convergence result of Algorithm 10.1. While recent work on the global guarantees of the algorithm has shown the linear behavior under certain statistical models, we emphasize that no closed-form expression of the convergence rate was provided. Our result in this chapter, on the other hand, does not make any assumption about the underlying model for $\boldsymbol{M}$ and Ω, and provides an exact expression of the asymptotic rate of convergence. Let us first introduce some critical concepts used in our derivation.

Definition 10.1. *Denote* $\bar{\Omega} = \{(i-1)n+j \mid (i,j) \in \Omega\}$. *The row selection matrix* $\boldsymbol{S}$ *is an* $s \times n^2$ *matrix obtained from a subset of rows corresponding to the elements in* $\bar{\Omega}$ *from the* $n^2 \times n^2$ *identity matrix* $\boldsymbol{I}_{n^2}$.

Definition 10.2. *The orthogonal projection onto the null space of* $\boldsymbol{M}$ *is defined by* $\boldsymbol{P}_{\boldsymbol{U}_\perp} = \boldsymbol{I}_n - \boldsymbol{U}\boldsymbol{U}^\top$.

Definition 10.3. *Let* $\boldsymbol{T}_{n^2}$ *be an* $n^2 \times n^2$ *matrix where the* (i,j)*th block of* $\boldsymbol{T}_{n^2}$ *is the* $n \times n$ *matrix* $\boldsymbol{e}_j\boldsymbol{e}_i^\top$ *for* $1 \leq i,j \leq n$. *Then* $\boldsymbol{T}_{n^2}$ *can be used to represent the transpose operator as follows:*

$$\operatorname{vec}(\boldsymbol{E}^\top) = \boldsymbol{T}_{n^2} \operatorname{vec}(\boldsymbol{E}) \quad \text{for any } \boldsymbol{E} \in \mathbb{R}^{n \times n}.$$

We are now in position to state our main result on the asymptotic linear convergence rate of Algorithm 10.1.

Theorem 10.1. *Denote* $\boldsymbol{P}_1 = \boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}$, $\boldsymbol{P}_2 = \frac{1}{2}\big(\boldsymbol{I}_{n^2} + \boldsymbol{T}_{n^2}\big)$, *and* $\boldsymbol{P} = \boldsymbol{P}_1\boldsymbol{P}_2$. *In addition, let*

$$\boldsymbol{H} = \boldsymbol{P}\Big(\boldsymbol{I}_{n^2} - \eta(\boldsymbol{M} \oplus \boldsymbol{M})(\boldsymbol{S}^\top \boldsymbol{S})\Big)\boldsymbol{P},$$

where $\boldsymbol{M} \oplus \boldsymbol{M} = \boldsymbol{M} \otimes \boldsymbol{I}_n + \boldsymbol{I}_n \otimes \boldsymbol{M}$ *is the Kronecker sum. Define the spectral radius of* $\boldsymbol{H}$, $\rho(\boldsymbol{H})$, *as the largest absolute value of the eigenvalues of* $\boldsymbol{H}$. *If* $\rho(\boldsymbol{H}) < 1$, *then Algorithm 10.1 produces a sequence of matrices* $\boldsymbol{X}^k \boldsymbol{X}^{k\top}$ *converging to* $\boldsymbol{M}$ *at an asymptotic linear rate* $\rho(\boldsymbol{H})$. *Formally, there exists a neighborhood* $\mathcal{N}(\boldsymbol{M})$ *of* $\boldsymbol{M}$ *such that for any* $\boldsymbol{X}^0 \boldsymbol{X}^{0\top} \in \mathcal{N}(\boldsymbol{M})$,

$$\left\| \boldsymbol{X}^k \boldsymbol{X}^{k\top} - \boldsymbol{M} \right\|_F \leq C \left\| \boldsymbol{X}^0 \boldsymbol{X}^{0\top} - \boldsymbol{M} \right\|_F \rho(\boldsymbol{H})^k, \tag{10.3}$$

for some numerical constant $C > 0$.

Remark 10.1. *Theorem 10.1 provides a closed-form expression of the asymptotic linear convergence rate of Algorithm 10.1, which only depends on* $\boldsymbol{M}$, Ω *and the choice of step-size* η. *We note that the condition for linear convergence,* $\rho(\boldsymbol{H}) < 1$, *is fully determined given* $\boldsymbol{M}$, Ω, *and* η. *It would be interesting to establish a connection between this condition and the standard statistical model for matrix completion. For instance, how the incoherence of* $\boldsymbol{M}$ *and the randomness of* Ω *would affect the spectral radius of* $\boldsymbol{H}$*? This exploration is left as future work.*

In our approach, the following lemma plays a pivotal role in the derivation of

Theorem 10.1, establishing the recursion on the error matrix $\boldsymbol{X}^k\boldsymbol{X}^{k^\top} - \boldsymbol{M}$:[3]

Lemma 10.1. *Let* $\boldsymbol{E}^k = \boldsymbol{X}^k\boldsymbol{X}^{k^\top} - \boldsymbol{M}$. *Then*

$$\boldsymbol{E}^{k+1} = \boldsymbol{E}^k - \eta\big(\mathcal{P}_\Omega(\boldsymbol{E}^k)\boldsymbol{M} + \boldsymbol{M}\mathcal{P}_\Omega(\boldsymbol{E}^k)\big) + \boldsymbol{\mathcal{O}}(\left\|\boldsymbol{E}^k\right\|_F^2).$$

Furthermore, denote $\boldsymbol{A} = \boldsymbol{I}_{n^2} - \eta(\boldsymbol{M} \oplus \boldsymbol{M})(\boldsymbol{S}^\top\boldsymbol{S})$ *and* $\boldsymbol{e}^k = \mathrm{vec}(\boldsymbol{E}^k)$, *the matrix recursion can be rewritten compactly as*

$$\boldsymbol{e}^{k+1} = \boldsymbol{A}\boldsymbol{e}^k + \boldsymbol{\mathcal{O}}(\left\|\boldsymbol{e}^k\right\|_2^2). \tag{10.4}$$

Remark 10.2. *Figure 10.1 illustrates the effectiveness of the proposed bound on the asymptotic rate of convergence given by Theorem 10.1. In Fig. 10.1, the low-rank solution matrix* $\boldsymbol{M}$ *is generated by taking the product of a* 20×3 *matrix* $\boldsymbol{X}$ *and its transpose, where* $\boldsymbol{X}$ *has i.i.d. normally distributed entries. The sampling set* Ω *is obtained by randomly selecting the entries of* $\boldsymbol{M}$ *based on a Bernoulli model with probability* 0.3. *Next, we run the economy-SVD on* $\boldsymbol{M}$ *to compute* $\boldsymbol{X}^* = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$. *The initialization* $\boldsymbol{X}^0$ *is obtained by adding i.i.d. normally distributed noise with standard deviation* $\sigma = 10^{-2}$ *to the entries of* $\boldsymbol{X}^*$. *Then we run Algorithm 10.1 with* $\boldsymbol{X}^0$, $\mathcal{P}_\Omega(\boldsymbol{M})$, *and* $\eta = 0.5/\left\|\boldsymbol{M}\right\|_2$. *It is noticeable from Fig. 10.1 that our theoretical bound* $\left\|\boldsymbol{e}^0\right\|_2 \rho(\boldsymbol{H})^k$ *given by the green line predicts successfully the rate of decrease in* $\left\|\boldsymbol{E}^k\right\|_F$, *running parallel to the blue line as soon as* $\left\|\boldsymbol{E}^k\right\|_F < 10^{-2}$. *As far as the approximations are concerned, we compare the changes in the error*

[3]We provide proofs of all the lemmas in the appendix at the end of this chapter.

modeled by $\boldsymbol{e}^{k+1} = \boldsymbol{A}\boldsymbol{e}^k$ *and the error modeled by* $\boldsymbol{e}^{k+1} = \boldsymbol{H}\boldsymbol{e}^k$. *While the former (represented by* $\left\|\boldsymbol{A}^k\boldsymbol{e}^0\right\|_2$ *in black) fails to approximate* $\left\|\boldsymbol{E}^k\right\|_F$ *for* $\left\|\boldsymbol{E}^k\right\|_F < 10^{-2}$, *the later (represented by* $\left\|\boldsymbol{H}^k\boldsymbol{e}^0\right\|_2$ *in red) matches* $\left\|\boldsymbol{E}^k\right\|_F$ *surprisingly well.*

In the rest of this section, we shall derive the proof of Theorem 10.1. First, we present a major challenge met by the traditional approach that uses (10.4) to characterize the convergence of the error towards zero. Next, we describe our proposed technique to overcome this difficulty. Finally, we show that our bounding technique recovers the exact rate of local convergence of Algorithm 10.1.

10.3.1 A Challenge of Establishing the Error Contraction

The stability of the nonlinear difference equation (10.4) is the key to analyze the convergence of Algorithm 10.1. In essence, linear convergence rate is obtained by the following lemma:

Lemma 10.2. *(Rephrased from the supplemental material of [212]) Let* $(a_n)_{n\in\mathbb{N}} \subset \mathbb{R}_+$ *be the sequence defined by*

$$a_{n+1} \leq \rho a_n + q a_n^2 \qquad \text{for } n = 0, 1, 2, \ldots,$$

where $0 < \rho < 1$ *and* $q \geq 0$. *Then* (a_n) *converges to* 0 *if and only if* $a_0 < \frac{1-\rho}{q}$. *A simple linear convergence bound can be derived for* $a_0 < \rho\frac{1-\rho}{q}$ *in the form of*

$$a_n \leq a_0 K \rho^n, \quad \text{for } K = \left(1 - \frac{a_0 q}{\rho(1-\rho)}\right)^{-1}.$$

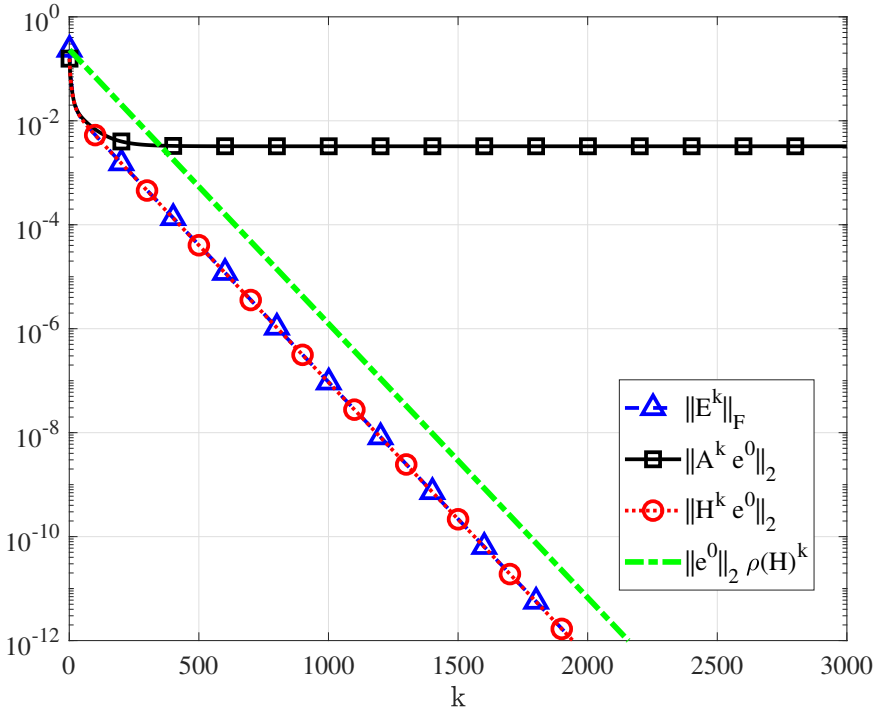


Figure 10.1: Linear convergence of gradient descent for matrix completion. The decrease in the norm of error matrix $\boldsymbol{E}^k$ through iterations is shown in the blue dashed line with triangle markers. The black solid line with square markers and the red dotted line with circle markers represent first-order approximations of the error using $\boldsymbol{A}$ and $\boldsymbol{H}$, respectively. Finally, the green dash-dot line is the theoretical bound (up to a constant) given by $\|\boldsymbol{e}^0\|_2 \rho(\boldsymbol{H})^k$. We use different markers, i.e., triangle versus circle, to better distinguish the blue line from the red line, respectively.

In order to apply Lemma 10.2 to (10.4), one natural way is to perform the eigen-decomposition $\boldsymbol{A} = \boldsymbol{Q_A}\boldsymbol{\Lambda_A}\boldsymbol{Q_A}^{-1}$, where $\boldsymbol{Q_A}$ is the square matrix whose columns are n^2 eigenvectors of $\boldsymbol{A}$, and $\boldsymbol{\Lambda_A}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues of $\boldsymbol{A}$. Then, left-multiplying both sides of (10.4) by $\boldsymbol{Q_A}^{-1}$ yields

$$\boldsymbol{Q_A}^{-1}\boldsymbol{e}^{k+1} = \boldsymbol{\Lambda_A}\boldsymbol{Q_A}^{-1}\boldsymbol{e}^{k} + \boldsymbol{\mathcal{O}}(\left\|\boldsymbol{e}^{k}\right\|_2^2),$$

where $\boldsymbol{Q_A}^{-1}$ does not affect the $\boldsymbol{\mathcal{O}}$ term since its norm is constant. Applying the triangle inequality[4] to the last equation leads to

$$\left\|\boldsymbol{Q_A}^{-1}\boldsymbol{e}^{k+1}\right\|_2 = \left\|\boldsymbol{\Lambda_A}\boldsymbol{Q_A}^{-1}\boldsymbol{e}^{k}\right\|_2 + \mathcal{O}(\left\|\boldsymbol{e}^{k}\right\|_2^2). \tag{10.5}$$

With the definition of the spectral radius of $\boldsymbol{A}$ using the spectral norm of $\boldsymbol{\Lambda}_A$, we have

$$\rho(\boldsymbol{A}) = \left\|\boldsymbol{\Lambda_A}\right\|_2 = \sup\left\{\frac{\left\|\boldsymbol{\Lambda_A}\tilde{\boldsymbol{e}}\right\|_2}{\left\|\tilde{\boldsymbol{e}}\right\|_2} : \tilde{\boldsymbol{e}} \in \mathbb{R}^{n^2}, \tilde{\boldsymbol{e}} \neq \boldsymbol{0}\right\}. \tag{10.6}$$

Now, using (10.6) and the fact that $\mathcal{O}(\left\|\boldsymbol{e}^{k}\right\|_2^2) = \mathcal{O}(\left\|\boldsymbol{Q_A}^{-1}\boldsymbol{e}^{k}\right\|_2^2)$, (10.5) can be

[4]Given $a = b + c$, by triangle inequality, we have $\|a\| \leq \|b\| + \|c\|$ and $\|a\| \geq \|b\| - \|c\|$ (since $b = a + (-c)$ and hence $\|b\| \leq \|a\| + \| - c\| = \|a\| + \|c\|$ or $\|a\| \geq \|b\| - \|c\|$). Consequently, we can write $|\|a\| - \|b\|| \leq \|c\|$ and hence $\|a\| = \|b\| + \mathcal{O}(\|c\|)$.

upper-bounded by

$$\left\|\boldsymbol{Q}_{\boldsymbol{A}}^{-1}\boldsymbol{e}^{k+1}\right\|_2 \leq \rho(\boldsymbol{A})\left\|\boldsymbol{Q}_{\boldsymbol{A}}^{-1}\boldsymbol{e}^{k}\right\|_2 + \mathcal{O}(\left\|\boldsymbol{Q}_{\boldsymbol{A}}^{-1}\boldsymbol{e}^{k}\right\|_2^2). \tag{10.7}$$

If $\rho(\boldsymbol{A}) < 1$, then by Lemma 10.2, the sequence $\left\|\boldsymbol{Q}_{\boldsymbol{A}}^{-1}\boldsymbol{e}^{k}\right\|_2$ converges to 0 linearly at rate $\rho(\boldsymbol{A})$. Unfortunately, one can verify that $\rho(\boldsymbol{A}) \geq 1$ by taking any vector $\boldsymbol{v} \in \mathbb{R}^{n^2}$ such that $v_i = 0$ for all $i \in \bar{\Omega}$. Since $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{v}$, 1 must be an eigenvalue of $\boldsymbol{A}$.

The failure of the aforementioned bounding technique is it overlooks the fact that $\boldsymbol{E}^k = \boldsymbol{X}^k{\boldsymbol{X}^k}^\top - \boldsymbol{M}$. By defining $\mathcal{E} = \{\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M} \mid \boldsymbol{X} \in \mathbb{R}^{n\times r}\}$ and $\tilde{\mathcal{E}}_{\boldsymbol{A}} = \{\boldsymbol{Q}_{\boldsymbol{A}}^{-1}\operatorname{vec}(\boldsymbol{E}) \mid \boldsymbol{E} \in \mathcal{E}\}$, a tighter bound on $\left\|\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{Q}_{\boldsymbol{A}}^{-1}\boldsymbol{e}^{k}\right\|_2 / \left\|\boldsymbol{Q}_{\boldsymbol{A}}^{-1}\boldsymbol{e}^{k}\right\|_2$ can be obtained by

$$\rho^{\mathcal{E}}(\boldsymbol{A},\delta) = \sup\left\{\frac{\|\boldsymbol{\Lambda}_{\boldsymbol{A}}\tilde{\boldsymbol{e}}\|_2}{\|\tilde{\boldsymbol{e}}\|_2} : \tilde{\boldsymbol{e}} \in \tilde{\mathcal{E}}_{\boldsymbol{A}}, \tilde{\boldsymbol{e}} \neq \boldsymbol{0}, \|\tilde{\boldsymbol{e}}\|_2 \leq \delta\right\}, \tag{10.8}$$

for some constant $\delta > 0$. Taking into account the structure of $\boldsymbol{E}^k$, one would expect $\rho^{\mathcal{E}}(\boldsymbol{A}) = \lim_{\delta\to 0}\rho^{\mathcal{E}}(\boldsymbol{A},\delta)$ is a more reliable estimate of the asymptotic rate of convergence for (10.4). Nonetheless, (10.8) is a non-trivial optimization problem that has no closed-form solution to the best of our knowledge.

10.3.2 Integrating Structural Constraints

To address the aforementioned issue, we propose to integrate the structural constraint on $\boldsymbol{E}^k$ into the recursion (10.4). As we shall show in the next subsection,

this integration enables the application of Lemma 10.2 to the new recursion in order to obtain a tight bound on the convergence rate. First, let us characterize the feasible set of error matrices $\mathcal{E}$ as follows:

Lemma 10.3. *$\boldsymbol{E} \in \mathcal{E}$ if and only if the following conditions hold simultaneously:*

(C1) $\mathcal{P}_r(\boldsymbol{M}+\boldsymbol{E}) = \boldsymbol{M}+\boldsymbol{E}$*, where $\mathcal{P}_r$ is the truncated singular value decomposition of order r [61].*

(C2) $\boldsymbol{E}^\top = \boldsymbol{E}$.

(C3) $\boldsymbol{v}^\top(\boldsymbol{M}+\boldsymbol{E})\boldsymbol{v} \geq 0$ *for all* $\boldsymbol{v} \in \mathbb{R}^n$.

Our strategy is to integrate three conditions in Lemma 10.3 into the linear operator $\boldsymbol{A}$ so that the resulting recursion will implicitly enforce $\boldsymbol{E}^k$ to remain in $\mathcal{E}$. Specifically, for condition (C1), we linearize $\mathcal{P}_r$ using the first-order perturbation analysis of the truncated singular value decomposition [211]. For condition (C2), we leverage the linearity of the transpose operator. Finally, while handling condition (C3) is non-trivial, it turns out that this condition can be ignored. In the following lemma, we introduce the linear projection that ensures the updated error $\boldsymbol{E}^k$ remains near $\mathcal{E}$.

Lemma 10.4. *Recall that $\boldsymbol{P}_1 = \boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}$, $\boldsymbol{P}_2 = \frac{1}{2}\big(\boldsymbol{I}_{n^2} + \boldsymbol{T}_{n^2}\big)$. Then, the following statements hold:*

1. *$\boldsymbol{P}_1$ corresponds to the orthogonal projection onto the tangent plane of the set of rank-r matrices at $\boldsymbol{M}$.*

2. *$\boldsymbol{P}_2$ corresponds to the orthogonal projection onto the space of symmetric matrices.*

3. *$\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ commute, and $\boldsymbol{P} = \boldsymbol{P}_1\boldsymbol{P}_2 = \boldsymbol{P}_2\boldsymbol{P}_1$ is also an orthogonal projection.*

4. *For any $\boldsymbol{E} \in \mathcal{E}$, $\operatorname{vec}(\boldsymbol{E}) = \boldsymbol{P}\operatorname{vec}(\boldsymbol{E}) + \boldsymbol{\mathcal{O}}(\|\boldsymbol{E}\|_F^2)$.*

By Lemma 10.4-4, we have $\boldsymbol{e}^k = \boldsymbol{P}\boldsymbol{e}^k + \boldsymbol{\mathcal{O}}(\left\|\boldsymbol{e}^k\right\|_2^2)$ for all k. Using this result with $k+1$ instead of k and replacing $\boldsymbol{e}^{k+1}$ from (10.4) into the first term on the RHS, we have

$$\boldsymbol{e}^{k+1} = \boldsymbol{P}\big(\boldsymbol{A}\boldsymbol{e}^k + \boldsymbol{\mathcal{O}}(\left\|\boldsymbol{e}^k\right\|_2^2)\big) + \boldsymbol{\mathcal{O}}(\left\|\boldsymbol{e}^{k+1}\right\|_2^2).$$

Substituting $\boldsymbol{e}^k = \boldsymbol{P}\boldsymbol{e}^k + \boldsymbol{\mathcal{O}}(\left\|\boldsymbol{e}^k\right\|_2^2)$ and using $\boldsymbol{e}^{k+1} = \boldsymbol{\mathcal{O}}(\left\|\boldsymbol{e}^k\right\|_2)$, we obtain

$$\boldsymbol{e}^{k+1} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}^k + \boldsymbol{\mathcal{O}}(\left\|\boldsymbol{e}^k\right\|_2^2). \tag{10.9}$$

It can be seen from Lemma 10.4-1 and Lemma 10.4-2 that the projection $\boldsymbol{P}$ enforces the error vector $\boldsymbol{e}^k$ to lie in the space under conditions (C1) and (C2) in Lemma 10.3. Now replacing the definition $\boldsymbol{H} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}$, (10.9) can be rewritten as

$$\boldsymbol{e}^{k+1} = \boldsymbol{H}\boldsymbol{e}^k + \boldsymbol{\mathcal{O}}(\left\|\boldsymbol{e}^k\right\|_2^2). \tag{10.10}$$

Similar to the derivation with $\boldsymbol{A}$, let $\boldsymbol{H} = \boldsymbol{Q}_{\boldsymbol{H}}\boldsymbol{\Lambda}_{\boldsymbol{H}}\boldsymbol{Q}_{\boldsymbol{H}}^{-1}$ be the eigendecomposition of $\boldsymbol{H}$ and define $\tilde{\boldsymbol{e}}^k = \boldsymbol{Q}_{\boldsymbol{H}}^{-1}\boldsymbol{e}^k$. Then, we have

$$\left\|\tilde{\boldsymbol{e}}^{k+1}\right\|_2 = \left\|\boldsymbol{\Lambda}_{\boldsymbol{H}}\tilde{\boldsymbol{e}}^k\right\|_2 + \mathcal{O}(\left\|\tilde{\boldsymbol{e}}^k\right\|_2^2). \tag{10.11}$$

In addition, denote $\tilde{\mathcal{E}}_{\boldsymbol{H}} = \{\boldsymbol{Q}_{\boldsymbol{H}}^{-1}\operatorname{vec}(\boldsymbol{E}) \mid \boldsymbol{E} \in \mathcal{E}\}$, we can define

$$\rho(\boldsymbol{H}) = \sup\left\{ \frac{\|\boldsymbol{\Lambda}_{\boldsymbol{H}}\tilde{\boldsymbol{e}}\|_2}{\|\tilde{\boldsymbol{e}}\|_2} : \tilde{\boldsymbol{e}} \in \mathbb{R}^{n^2}, \tilde{\boldsymbol{e}} \neq \boldsymbol{0} \right\} \text{ and} \tag{10.12}$$

$$\rho^{\mathcal{E}}(\boldsymbol{H}, \delta) = \sup\left\{ \frac{\|\boldsymbol{\Lambda}_{\boldsymbol{H}}\tilde{\boldsymbol{e}}\|_2}{\|\tilde{\boldsymbol{e}}\|_2} : \tilde{\boldsymbol{e}} \in \tilde{\mathcal{E}}_{\boldsymbol{H}}, \tilde{\boldsymbol{e}} \neq \boldsymbol{0}, \|\tilde{\boldsymbol{e}}\|_2 \leq \delta \right\}. \tag{10.13}$$

Since (10.4) and (10.10) are two different systems that describes the same dynamic for $\boldsymbol{E}^k \in \mathcal{E}$, one would expect they share the same asymptotic behavior. In particular, their linear rates of convergence should agree when the constraint $\boldsymbol{E}^k \in \mathcal{E}$ is considered.

Lemma 10.5. *Let* $\rho^{\mathcal{E}}(\boldsymbol{H}) = \lim_{\delta \to 0} \rho^{\mathcal{E}}(\boldsymbol{H}, \delta)$. *Then,*

$$\rho^{\mathcal{E}}(\boldsymbol{H}) = \rho^{\mathcal{E}}(\boldsymbol{A}).$$

While using $\boldsymbol{H}$ instead of $\boldsymbol{A}$ preserves the system dynamic over $\mathcal{E}$, it provides updates of the error that ensure that it remains in $\mathcal{E}$. Consequently, we can ignore the constraints that are implicitly satisfied in our analysis when using $\boldsymbol{H}$.

10.3.3 Asymptotic Bound on the Linear Convergence Rate

We have seen in Subsection 10.3.1 that applying Lemma 10.2 to (10.4) fails to estimate the convergence rate due to the gap between $\rho^{\mathcal{E}}(\boldsymbol{A})$ and $\rho(\boldsymbol{A})$. In this subsection, we show that integrating the structural constraint helps eliminating the gap between $\rho^{\mathcal{E}}(\boldsymbol{H})$ and $\rho(\boldsymbol{H})$ (even when condition (C3) is omitted). Therefore,

applying Lemma 10.2 to (10.11) yields $\rho(\boldsymbol{H})$ as a tight bound on the convergence rate. To that end, our goal is to prove the following lemma:

Lemma 10.6. *As δ approaches 0, we have $\rho(\boldsymbol{H})-\rho^{\mathcal{E}}(\boldsymbol{H},\delta)=\mathcal{O}(\delta)$. Consequently, it holds that $\rho(\boldsymbol{H})=\rho^{\mathcal{E}}(\boldsymbol{H})$.*

Let us briefly present the key ideas and lemmas we use to prove Lemma 10.6. Our proof relies on two critical considerations: (i) $\rho^{\mathcal{E}}(\boldsymbol{H},\delta)\leq\rho(\boldsymbol{H})$, (ii) there exists a maximizer $\tilde{\boldsymbol{e}}^{\star}$ of the supremum in (10.12) such that the distance from $\tilde{\boldsymbol{e}}^{\star}$ to $\tilde{\mathcal{E}}_{\boldsymbol{H}}$ is $\mathcal{O}(\delta^2)$. While (i) is trivial from (10.12) and (10.13), (ii) is proven by introducing $\mathcal{F}_{\delta}$ as a surrogate for the set $\mathcal{E}$ as follows:

Lemma 10.7. *Denote the eigenvector of $\boldsymbol{H}$ corresponding to the largest (in magnitude) eigenvalue by $\boldsymbol{q}_1$. Define $\boldsymbol{G}$ as the $n\times n$ matrix satisfying $\mathrm{vec}(\boldsymbol{G})=\delta\boldsymbol{q}_1$. Let $\mathcal{F}_{\delta}$ be the set of $n\times n$ matrices satisfying the following conditions: (i) $\|\boldsymbol{F}\|_F\leq 2\delta$; (ii) $\boldsymbol{F}^{\top}=\boldsymbol{F}$; (iii) $\|\boldsymbol{P}_{\boldsymbol{U}_{\perp}}\boldsymbol{F}\boldsymbol{P}_{\boldsymbol{U}_{\perp}}\|_F\leq\frac{2}{\lambda_r}\delta^2$; and (iv) $\boldsymbol{v}^{\top}(\boldsymbol{M}+\boldsymbol{F})\boldsymbol{v}\geq 0$ for all $\boldsymbol{v}\in\mathbb{R}^n$. Then, there exists $\boldsymbol{F}\in\mathcal{F}_{\delta}$ satisfying*

$$\|\boldsymbol{F}-\boldsymbol{G}\|_F=\mathcal{O}(\delta^2).$$

Lemma 10.8. *For any $\boldsymbol{F}\in\mathcal{F}_{\delta}$, there exists $\boldsymbol{E}\in\mathcal{E}$ satisfying*

$$\|\boldsymbol{E}-\boldsymbol{F}\|_F=\mathcal{O}(\delta^2).$$

From (i) and (ii), it follows that the difference between $\rho^{\mathcal{E}}(\boldsymbol{H},\delta)$ and $\rho(\boldsymbol{H})$ is $\mathcal{O}(\delta)$. Thus, $\rho(\boldsymbol{H})=\rho^{\mathcal{E}}(\boldsymbol{H})$ when taking the limit of $\rho^{\mathcal{E}}(\boldsymbol{H},\delta)$ as $\delta\to 0$. Our derivation

of Theorem 10.1 is completed by directly applying Lemma 10.2 to (10.11).

10.4 Conclusion and Future work

We presented a framework for analyzing the convergence of the existing gradient descent approach for low-rank matrix completion. In our analysis, we restricted our focus to the symmetric matrix completion case. We proved that the algorithm converges linearly. Different to other approaches, we made no assumption on the rank of the matrix or fraction of available entries. Instead, we derived an expression for the linear convergence rate via the spectral norm of a closed-form matrix. As future work, using random matrix theory, the closed-form expression for the convergence rate can be further related to the rank, the number of available entries, and the matrix dimensions. Additionally, this work can be extended to the non-symmetric case.

10.5 Appendix

10.5.1 Proof of Lemma 10.1

Recall the gradient descent update in Algorithm 10.1:

$$\begin{aligned} \boldsymbol{X}^{k+1} &= \boldsymbol{X}^k - \eta \mathcal{P}_\Omega\big(\boldsymbol{X}^k {\boldsymbol{X}^k}^\top - \boldsymbol{M}\big)\boldsymbol{X}^k \\ &= (\boldsymbol{I}_n - \eta \mathcal{P}_\Omega(\boldsymbol{E}^k))\boldsymbol{X}^k. \end{aligned} \tag{10.14}$$

Substituting (10.14) into the definition of $\boldsymbol{E}^{k+1}$, we have

$$\begin{aligned}\boldsymbol{E}^{k+1} &= \boldsymbol{X}^{k+1}\boldsymbol{X}^{k+1^\top} - \boldsymbol{M} \\ &= \big(\boldsymbol{I}_n - \eta\mathcal{P}_\Omega(\boldsymbol{E}^k)\big)\boldsymbol{X}^k\boldsymbol{X}^{k^\top}\big(\boldsymbol{I}_n - \eta\mathcal{P}_\Omega(\boldsymbol{E}^k)\big)^\top - \boldsymbol{M}.\end{aligned}$$

From the fact that $\boldsymbol{E}^k$ is symmetric and Ω is a symmetric sampling, the last equation can be further expanded as

$$\begin{aligned}\boldsymbol{E}^{k+1} = \boldsymbol{X}^k\boldsymbol{X}^{k^\top} &- \eta\mathcal{P}_\Omega(\boldsymbol{E}^k)\boldsymbol{X}^k\boldsymbol{X}^{k^\top} \\ &- \eta\boldsymbol{X}^k\boldsymbol{X}^{k^\top}\mathcal{P}_\Omega(\boldsymbol{E}^k) + \eta^2\mathcal{P}_\Omega(\boldsymbol{E}^k)\boldsymbol{X}^k\boldsymbol{X}^{k^\top}\mathcal{P}_\Omega(\boldsymbol{E}^k) - \boldsymbol{M}. \quad (10.15)\end{aligned}$$

Since $\boldsymbol{X}^k\boldsymbol{X}^{k^\top} = \boldsymbol{M} + \boldsymbol{E}^k$, (10.15) is equivalent to

$$\begin{aligned}\boldsymbol{E}^{k+1} = \boldsymbol{E}^k &- \eta\big(\mathcal{P}_\Omega(\boldsymbol{E}^k)\boldsymbol{M} + \boldsymbol{M}\mathcal{P}_\Omega(\boldsymbol{E}^k)\big) - \eta\big(\mathcal{P}_\Omega(\boldsymbol{E}^k)\boldsymbol{E}^k + \boldsymbol{E}^k\mathcal{P}_\Omega(\boldsymbol{E}^k)\big) \\ &+ \eta^2\mathcal{P}_\Omega(\boldsymbol{E}^k)\boldsymbol{M}\mathcal{P}_\Omega(\boldsymbol{E}^k) + \eta^2\mathcal{P}_\Omega(\boldsymbol{E}^k)\boldsymbol{E}^k\mathcal{P}_\Omega(\boldsymbol{E}^k). \quad (10.16)\end{aligned}$$

Note that $\big\|\mathcal{P}_\Omega(\boldsymbol{E}^k)\big\|_F \le \big\|\boldsymbol{E}^k\big\|_F$. Hence, collecting terms that are of second order and higher, with respect to $\big\|\boldsymbol{E}^k\big\|_F$, on the RHS of (10.16) yields

$$\boldsymbol{E}^{k+1} = \boldsymbol{E}^k - \eta\big(\mathcal{P}_\Omega(\boldsymbol{E}^k)\boldsymbol{M} + \boldsymbol{M}\mathcal{P}_\Omega(\boldsymbol{E}^k)\big) + \boldsymbol{\mathcal{O}}(\big\|\boldsymbol{E}^k\big\|_F^2).$$

Now by Definition 7.2, it is easy to verify that

$$\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{I}_{n^2} \quad \text{and} \quad \mathrm{vec}\big(\mathcal{P}_\Omega(\boldsymbol{E}^k)\big) = \boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{e}^k.$$

Using the property $\operatorname{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^\top \otimes \boldsymbol{A})\operatorname{vec}(\boldsymbol{B})$, (5) can be vectorized as follows:

$$\boldsymbol{e}^{k+1} = \boldsymbol{e}^k - \eta(\boldsymbol{M} \otimes \boldsymbol{I}_n)\operatorname{vec}\big(\mathcal{P}_\Omega(\boldsymbol{E}^k)\big) - \eta(\boldsymbol{I}_n \otimes \boldsymbol{M})\operatorname{vec}\big(\mathcal{P}_\Omega(\boldsymbol{E}^k)\big) + \boldsymbol{\mathcal{O}}(\big\|\boldsymbol{e}^k\big\|_2^2).$$

The last equation can be reorganized as

$$\boldsymbol{e}^{k+1} = \Big(\boldsymbol{I}_{n^2} - \eta(\boldsymbol{M} \oplus \boldsymbol{M})(\boldsymbol{S}^\top \boldsymbol{S})\Big)\boldsymbol{e}^k + \boldsymbol{\mathcal{O}}(\big\|\boldsymbol{e}^k\big\|_2^2).$$

10.5.2 Proof of Lemma 10.3

($\Rightarrow$) Suppose $\boldsymbol{E} \in \mathcal{E}$. Then for (C1), i.e., $\boldsymbol{E}^\top = \boldsymbol{E}$, $\boldsymbol{E} = \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}$ is symmetric since both $\boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{M}$ are symmetric. For (C2), i.e., $\mathcal{P}_r(\boldsymbol{M} + \boldsymbol{E}) = \boldsymbol{M} + \boldsymbol{E}$, stems from the fact $\boldsymbol{M} + \boldsymbol{E} = \boldsymbol{X}\boldsymbol{X}^\top$ has rank no greater than r for $\boldsymbol{X} \in \mathbb{R}^{n\times r}$. Finally, for any $\boldsymbol{v} \in \mathbb{R}^n$, we have

$$\boldsymbol{v}^\top(\boldsymbol{M} + \boldsymbol{E})\boldsymbol{v} = \boldsymbol{v}^\top(\boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{v} = \big\|\boldsymbol{X}^\top\boldsymbol{v}\big\|_2^2 \geq 0.$$

($\Leftarrow$) From conditions (C1) and (C3), $\boldsymbol{M} + \boldsymbol{E}$ is a PSD matrix. In addition, $\mathcal{P}_r(\boldsymbol{M} + \boldsymbol{E}) = \boldsymbol{M} + \boldsymbol{E}$ implies $\boldsymbol{M} + \boldsymbol{E}$ must have rank no greater r. Since any PSD matrix $\boldsymbol{A}$ with rank less than or equal to r can be factorized as $\boldsymbol{A} = \boldsymbol{Y}\boldsymbol{Y}^\top$ for some $\boldsymbol{Y} \in \mathbb{R}^{n\times r}$, we conclude that $\boldsymbol{E} \in \mathcal{E}$.

10.5.3 Proof of Lemma 10.4

First, recall that any matrix $\boldsymbol{\Pi} \in \mathbb{R}^{n^2 \times n^2}$ is an orthogonal projection if and only if $\boldsymbol{\Pi}^2 = \boldsymbol{\Pi}$ and $\boldsymbol{\Pi} = \boldsymbol{\Pi}^\top$. Since $\boldsymbol{P}_{\boldsymbol{U}_\perp}^\top = \boldsymbol{P}_{\boldsymbol{U}_\perp}$, we have

$$\begin{aligned}
\boldsymbol{P}_1^\top &= \left(\boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}\right)^\top \\
&= \boldsymbol{I}_{n^2}^\top - \boldsymbol{P}_{\boldsymbol{U}_\perp}^\top \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}^\top \\
&= \boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp} = \boldsymbol{P}_1.
\end{aligned}$$

In addition, since $\boldsymbol{P}_{\boldsymbol{U}_\perp}^2 = \boldsymbol{P}_{\boldsymbol{U}_\perp}$, we have

$$\begin{aligned}
\boldsymbol{P}_1^2 &= (\boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})(\boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})^\top \\
&= \boldsymbol{I}_{n^2}^2 - 2\boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp} + (\boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})^2 \\
&= \boldsymbol{I}_{n^2} - 2\boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp} + (\boldsymbol{P}_{\boldsymbol{U}_\perp}^2 \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}^2) \\
&= \boldsymbol{I}_{n^2} - 2\boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp} + \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp} \\
&= \boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp} = \boldsymbol{P}_1.
\end{aligned}$$

Second, using the fact that $\boldsymbol{T}_{n^2}^2 = \boldsymbol{I}_{n^2}$ and $\boldsymbol{T}_{n^2}$ is symmetric, we can derive similar result:

$$\boldsymbol{P}_2^\top = \left(\frac{\boldsymbol{I}_{n^2} + \boldsymbol{T}_{n^2}}{2}\right)^\top = \frac{\boldsymbol{I}_{n^2} + \boldsymbol{T}_{n^2}}{2} = \boldsymbol{P}_2,$$

and

$$\begin{aligned}\boldsymbol{P}_2^2 &= \frac{(\boldsymbol{I}_{n^2} + \boldsymbol{T}_{n^2})^2}{4} \\ &= \frac{\boldsymbol{I}_{n^2} + 2\boldsymbol{T}_{n^2} + \boldsymbol{T}_{n^2}^2}{4} \\ &= \frac{2\boldsymbol{I}_{n^2} + 2\boldsymbol{T}_{n^2}}{4} \\ &= \frac{\boldsymbol{I}_{n^2} + \boldsymbol{T}_{n^2}}{2} = \boldsymbol{P}_2.\end{aligned}$$

Third, we observe that $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are the vectorized version of the linear operators

$$\boldsymbol{\Pi}_1(\boldsymbol{E}) = \boldsymbol{E} - \boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{E}\boldsymbol{P}_{\boldsymbol{U}_\perp}$$

and

$$\boldsymbol{\Pi}_2(\boldsymbol{E}) = \frac{1}{2}(\boldsymbol{E} + \boldsymbol{E}^\top),$$

respectively, for any $\boldsymbol{E} \in \mathbb{R}^{n\times n}$. Hence, in order to prove that $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ commute, it is sufficient to show that operators $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$ commute. Indeed, we have

$$\begin{aligned}\boldsymbol{\Pi}_2\boldsymbol{\Pi}_1(\boldsymbol{E}) &= \frac{1}{2}\big((\boldsymbol{E} - \boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{E}\boldsymbol{P}_{\boldsymbol{U}_\perp}) + (\boldsymbol{E} - \boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{E}\boldsymbol{P}_{\boldsymbol{U}_\perp})^\top\big) \\ &= \frac{1}{2}(\boldsymbol{E} + \boldsymbol{E}^\top) - \boldsymbol{P}_{\boldsymbol{U}_\perp}\frac{1}{2}(\boldsymbol{E} + \boldsymbol{E}^\top)\boldsymbol{P}_{\boldsymbol{U}_\perp} \\ &= \boldsymbol{\Pi}_1\boldsymbol{\Pi}_2(\boldsymbol{E}).\end{aligned}$$

This implies $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$ commute. Since $\boldsymbol{P}$ is the product of two commuting orthogonal projections, it is also an orthogonal projection.

Finally, let us restrict $\boldsymbol{E}$ to belong to $\mathcal{E}$ and denote $\boldsymbol{e} = \mathrm{vec}(\boldsymbol{E})$. Using Theorem 3 in [211], we have

$$\mathcal{P}_r(\boldsymbol{M} + \boldsymbol{E}) = \boldsymbol{M} + \boldsymbol{E} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{E} \boldsymbol{P}_{\boldsymbol{U}_\perp} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{E}\|_F^2). \tag{10.17}$$

Since $\mathcal{P}_r(\boldsymbol{M} + \boldsymbol{E}) = \boldsymbol{M} + \boldsymbol{E}$, it follows from (10.17) that

$$\boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{E} \boldsymbol{P}_{\boldsymbol{U}_\perp} = \boldsymbol{\mathcal{O}}(\|\boldsymbol{E}\|_F^2).$$

Vectorizing the last equation, we obtain

$$(\boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \boldsymbol{e} = \boldsymbol{\mathcal{O}}(\|\boldsymbol{E}\|_F^2). \tag{10.18}$$

On the other hand, since $\boldsymbol{E}$ is symmetric,

$$\boldsymbol{e} = \boldsymbol{T}_{n^2} \boldsymbol{e} = \Big(\frac{\boldsymbol{I}_{n^2} + \boldsymbol{T}_{n^2}}{2} \Big) \boldsymbol{e}. \tag{10.19}$$

From (10.18) and (10.19), we have

$$\begin{aligned} \boldsymbol{e} &= (\boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \boldsymbol{e} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{E}\|_F^2) \\ &= (\boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \Big(\frac{\boldsymbol{I}_{n^2} + \boldsymbol{T}_{n^2}}{2} \Big) \boldsymbol{e} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{E}\|_F^2). \end{aligned} \tag{10.20}$$

Substituting

$$\boldsymbol{P} = \boldsymbol{P}_1\boldsymbol{P}_2 = (\boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})\Big(\frac{\boldsymbol{I}_{n^2} + \boldsymbol{T}_{n^2}}{2}\Big)$$

into (10.20) completes our proof of the lemma.

10.5.4 Proof of Lemma 10.5

Let $\tilde{\mathcal{E}} = \{\mathrm{vec}(\boldsymbol{E}) \mid \boldsymbol{E} \in \mathcal{E}\}$. Recall that for any $\boldsymbol{e} \in \tilde{\mathcal{E}}$,

$$\boldsymbol{e} = \boldsymbol{P}\boldsymbol{e} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{e}\|_2^2).$$

Therefore, by the triangle inequality, we obtain

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{e}\|_2 &= \big\|\boldsymbol{A}\big(\boldsymbol{P}\boldsymbol{e} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{e}\|_2^2)\big)\big\| \\ &\leq \|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}\|_2 + \big\|\boldsymbol{A}\boldsymbol{\mathcal{O}}(\|\boldsymbol{e}\|_2^2)\big\|_2 .\end{aligned}$$

Since the second term on the RHS of the last inequality is $\mathcal{O}(\|\boldsymbol{e}\|_2^2)$, it is also $\mathcal{O}(\delta^2)$ for any $\boldsymbol{e} \in \tilde{\mathcal{E}}$ such that $\|\boldsymbol{e}\|_2 \leq \delta$. In other words,

$$\|\boldsymbol{A}\boldsymbol{e}\|_2 \leq \|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}\|_2 + \mathcal{O}(\delta^2). \tag{10.21}$$

Similarly, we also have,

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{e}\|_2 &\geq \|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}\|_2 - \left\|\boldsymbol{A}\boldsymbol{\mathcal{O}}(\|\boldsymbol{e}\|_2^2)\right\|_2 \\ &= \|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}\|_2 - \mathcal{O}(\delta^2). \end{aligned} \tag{10.22}$$

From (10.21) and (10.22), it follows that

$$\frac{\|\boldsymbol{A}\boldsymbol{e}\|_2}{\|\boldsymbol{e}\|_2} = \frac{\|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}\|_2}{\|\boldsymbol{e}\|_2} + \mathcal{O}(\delta). \tag{10.23}$$

Taking the limit of the supremum of (10.23) as $\delta \to 0$ yields

$$\begin{aligned}\rho^{\mathcal{E}}(\boldsymbol{A}) &= \lim_{\delta\to 0} \sup_{\substack{\boldsymbol{e}\in\tilde{\mathcal{E}} \\ \boldsymbol{e}\neq 0 \\ \|\boldsymbol{e}\|_2\leq\delta}} \frac{\|\boldsymbol{A}\boldsymbol{e}\|_2}{\|\boldsymbol{e}\|_2} \\ &= \lim_{\delta\to 0} \sup_{\substack{\boldsymbol{e}\in\tilde{\mathcal{E}} \\ \boldsymbol{e}\neq 0 \\ \|\boldsymbol{e}\|_2\leq\delta}} \frac{\|\boldsymbol{A}\boldsymbol{P}\boldsymbol{e}\|_2}{\|\boldsymbol{e}\|_2} = \rho^{\mathcal{E}}(\boldsymbol{A}\boldsymbol{P}). \end{aligned} \tag{10.24}$$

Now following similar argument in Lemma 10.6, we have

$$\begin{cases} \rho^{\mathcal{E}}(\boldsymbol{A}\boldsymbol{P}) = \rho(\boldsymbol{A}\boldsymbol{P}), \\ \rho^{\mathcal{E}}(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}) = \rho(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}). \end{cases} \tag{10.25}$$

Given (10.24) and (10.25), it remains to show that $\rho(\boldsymbol{AP}) = \rho(\boldsymbol{PAP})$. Indeed, using Gelfand's formula [77], we have

$$\rho(\boldsymbol{AP}) = \lim_{k\to\infty} \left\|(\boldsymbol{AP})^k\right\|_2^{1/k}$$
$$\text{and } \rho(\boldsymbol{PAP}) = \lim_{k\to\infty} \left\|(\boldsymbol{PAP})^k\right\|_2^{1/k}.$$

By the property of operator norms,

$$\left\|(\boldsymbol{AP})^k\right\|_2 = \left\|\boldsymbol{A}(\boldsymbol{PAP})^{k-1}\right\|_2 \leq \|\boldsymbol{A}\|_2 \left\|(\boldsymbol{PAP})^{k-1}\right\|_2.$$

Thus,

$$\left\|(\boldsymbol{AP})^k\right\|_2^{1/k} \leq \|\boldsymbol{A}\|_2^{1/k} \left(\left\|(\boldsymbol{PAP})^{k-1}\right\|_2^{1/(k-1)}\right)^{(k-1)/k}.$$

Taking the limit of both sides of the last inequality as $k \to \infty$ yields $\rho(\boldsymbol{AP}) \leq \rho(\boldsymbol{PAP})$. Similarly, since

$$\left\|(\boldsymbol{PAP})^k\right\|_2 = \left\|\boldsymbol{P}(\boldsymbol{AP})^k\right\|_2 \leq \left\|(\boldsymbol{AP})^k\right\|_2,$$

we also obtain $\rho(\boldsymbol{PAP}) \leq \rho(\boldsymbol{AP})$. This concludes our proof of the lemma.

10.5.5 Proof of Lemma 10.6

Without loss of generality, assume λ_1 is the eigenvalue with largest magnitude, i.e., $|\lambda_1| = \rho(\boldsymbol{H})$. By the definition of $\boldsymbol{G}$, we have $\|\boldsymbol{G}\|_F = \delta$. Since $\boldsymbol{H}\,\mathrm{vec}(\boldsymbol{G}) = \lambda_1\,\mathrm{vec}(\boldsymbol{G})$ and $\boldsymbol{H} = \boldsymbol{Q_H}\boldsymbol{\Lambda_H}\boldsymbol{Q_H^{-1}}$, it follows that

$$\boldsymbol{Q_H}\boldsymbol{\Lambda_H}\boldsymbol{Q_H^{-1}}\,\mathrm{vec}(\boldsymbol{G}) = \lambda_1\,\mathrm{vec}(\boldsymbol{G}). \tag{10.26}$$

Multiplying both sides of (10.26) by $\boldsymbol{Q_H^{-1}}$, we obtain

$$\boldsymbol{\Lambda_H}\boldsymbol{Q_H^{-1}}\,\mathrm{vec}(\boldsymbol{G}) = \lambda_1\boldsymbol{Q_H^{-1}}\,\mathrm{vec}(\boldsymbol{G}).$$

Taking the $L2$-norm and and reorganizing the equation yields

$$\frac{\left\|\boldsymbol{\Lambda_H}\boldsymbol{Q_H^{-1}}\,\mathrm{vec}(\boldsymbol{G})\right\|_2}{\left\|\boldsymbol{Q_H^{-1}}\,\mathrm{vec}(\boldsymbol{G})\right\|_2} = |\lambda_1| = \rho(\boldsymbol{H}). \tag{10.27}$$

Therefore, $\boldsymbol{G}$ leads to a solution of the supremum in (10.12). We now prove that $\boldsymbol{G}$ is symmetric and $(\boldsymbol{P_{U_\perp}} \otimes \boldsymbol{P_{U_\perp}})\,\mathrm{vec}(\boldsymbol{G}) = \boldsymbol{0}$. First, since $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and $\boldsymbol{P} = \boldsymbol{P}_1\boldsymbol{P}_2$

are orthogonal projections, we have

$$\begin{aligned}\boldsymbol{P}_2\boldsymbol{H} &= \boldsymbol{P}_2\boldsymbol{P}\boldsymbol{A}\boldsymbol{P} \\ &= \boldsymbol{P}_2\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{P} \\ &= \boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{P} \\ &= \boldsymbol{P}_1\boldsymbol{P}_2\boldsymbol{A}\boldsymbol{P} \\ &= \boldsymbol{P}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{H}.\end{aligned}$$

Thus,

$$\begin{aligned}\lambda_1 \operatorname{vec}(\boldsymbol{G}) &= \boldsymbol{H} \operatorname{vec}(\boldsymbol{G}) \\ &= \boldsymbol{P}_2\boldsymbol{H} \operatorname{vec}(\boldsymbol{G}) \\ &= \lambda_1 \boldsymbol{P}_2 \operatorname{vec}(\boldsymbol{G}).\end{aligned} \tag{10.28}$$

Substituting $\boldsymbol{P}_2 = \frac{1}{2}\big(\boldsymbol{I}_{n^2} + \boldsymbol{T}_{n^2}\big)$ into (10.28) yields

$$\operatorname{vec}(\boldsymbol{G}^\top) = \boldsymbol{T}_{n^2} \operatorname{vec}(\boldsymbol{G}) \quad \text{or} \quad \boldsymbol{G} = \boldsymbol{G}^\top.$$

Second, since $\boldsymbol{P}_1\boldsymbol{H} = \boldsymbol{H}$, we obtain

$$\begin{aligned}\lambda_1 \operatorname{vec}(\boldsymbol{G}) &= \boldsymbol{H} \operatorname{vec}(\boldsymbol{G}) \\ &= \boldsymbol{P}_1\boldsymbol{H} \operatorname{vec}(\boldsymbol{G}) \\ &= \lambda_1 \boldsymbol{P}_1 \operatorname{vec}(\boldsymbol{G}).\end{aligned} \tag{10.29}$$

Substituting $\boldsymbol{P}_1 = \boldsymbol{I}_{n^2} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}$ into (10.29) yields

$$(\boldsymbol{P}_{\boldsymbol{U}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \operatorname{vec}(\boldsymbol{G}) = \boldsymbol{0} \quad \text{or} \quad \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{G} \boldsymbol{P}_{\boldsymbol{U}_\perp} = \boldsymbol{0}.$$

Since $\|\boldsymbol{E} - \boldsymbol{G}\|_F \leq \|\boldsymbol{E} - \boldsymbol{F}\|_F + \|\boldsymbol{F} - \boldsymbol{G}\|_F$ (by the triangle inequality), Lemmas 10.7 and 6.4 imply the existence of $\boldsymbol{E} \in \mathcal{E}$ such that $\|\boldsymbol{E} - \boldsymbol{G}\|_F = \mathcal{O}(\delta^2)$. Denote $\tilde{\boldsymbol{e}} = \boldsymbol{Q}_{\boldsymbol{H}}^{-1} \operatorname{vec}(\boldsymbol{E}) \in \tilde{\mathcal{E}}_{\boldsymbol{H}}$, we have

$$\begin{aligned}
\boldsymbol{\Lambda}_{\boldsymbol{H}} \tilde{\boldsymbol{e}} &= \lambda_1 \tilde{\boldsymbol{e}} - (\lambda_1 \boldsymbol{I}_{n^2} - \boldsymbol{\Lambda}_{\boldsymbol{H}}) \tilde{\boldsymbol{e}} \\
&= \lambda_1 \tilde{\boldsymbol{e}} - (\lambda_1 \boldsymbol{I}_{n^2} - \boldsymbol{\Lambda}_{\boldsymbol{H}}) \boldsymbol{Q}_{\boldsymbol{H}}^{-1} \operatorname{vec}(\boldsymbol{E}) \\
&= \lambda_1 \tilde{\boldsymbol{e}} - (\lambda_1 \boldsymbol{I}_{n^2} - \boldsymbol{\Lambda}_{\boldsymbol{H}}) \boldsymbol{Q}_{\boldsymbol{H}}^{-1} \operatorname{vec}(\boldsymbol{E} - \boldsymbol{G}).
\end{aligned}$$

where the last equality stems from the fact that $\lambda_1 \boldsymbol{Q}_{\boldsymbol{H}}^{-1} \operatorname{vec}(\boldsymbol{G}) = \boldsymbol{\Lambda}_{\boldsymbol{H}} \boldsymbol{Q}_{\boldsymbol{H}}^{-1} \operatorname{vec}(\boldsymbol{G})$. Next, using the triangle inequality, we obtain

$$\begin{aligned}
\|\boldsymbol{\Lambda}_{\boldsymbol{H}} \tilde{\boldsymbol{e}}\|_2 &\geq \|\lambda_1 \tilde{\boldsymbol{e}}\|_2 - \left\|(\lambda_1 \boldsymbol{I}_{n^2} - \boldsymbol{\Lambda}_{\boldsymbol{H}}) \boldsymbol{Q}_{\boldsymbol{H}}^{-1} \operatorname{vec}(\boldsymbol{E} - \boldsymbol{G})\right\|_2 \\
&\geq \rho(\boldsymbol{H}) \|\tilde{\boldsymbol{e}}\|_2 - \|\lambda_1 \boldsymbol{I}_{n^2} - \boldsymbol{\Lambda}_{\boldsymbol{H}}\|_2 \left\|\boldsymbol{Q}_{\boldsymbol{H}}^{-1}\right\|_2 \|\operatorname{vec}(\boldsymbol{E} - \boldsymbol{G})\|_2 .
\end{aligned}$$

Dividing both sides by $\|\tilde{\boldsymbol{e}}\|_2$ yields

$$\frac{\|\boldsymbol{\Lambda}_{\boldsymbol{H}} \tilde{\boldsymbol{e}}\|_2}{\|\tilde{\boldsymbol{e}}\|_2} \geq \rho(\boldsymbol{H}) - \frac{\|\lambda_1 \boldsymbol{I}_{n^2} - \boldsymbol{\Lambda}_{\boldsymbol{H}}\|_2 \left\|\boldsymbol{Q}_{\boldsymbol{H}}^{-1}\right\|_2 \|\operatorname{vec}(\boldsymbol{E} - \boldsymbol{G})\|_2}{\|\tilde{\boldsymbol{e}}\|_2}. \tag{10.30}$$

Since $\|\boldsymbol{E}-\boldsymbol{G}\|_F = \mathcal{O}(\delta^2)$, (10.30) can be rewritten as

$$\frac{\|\boldsymbol{\Lambda}_{\boldsymbol{H}}\tilde{\boldsymbol{e}}\|_2}{\|\tilde{\boldsymbol{e}}\|_2} \geq \rho(\boldsymbol{H}) - \mathcal{O}(\delta^2). \tag{10.31}$$

On the other hand, for any $\tilde{\boldsymbol{e}} \in \tilde{\mathcal{E}}_{\boldsymbol{H}}$, we also have

$$\frac{\|\boldsymbol{\Lambda}_{\boldsymbol{H}}\tilde{\boldsymbol{e}}\|_2}{\|\tilde{\boldsymbol{e}}\|_2} \leq \rho^{\mathcal{E}}(\boldsymbol{H},\delta) \leq \rho(\boldsymbol{H}). \tag{10.32}$$

Combining (10.31) and (10.32) yields $\rho(\boldsymbol{H}) - \rho^{\mathcal{E}}(\boldsymbol{H},\delta) = \mathcal{O}(\delta)$.

10.5.6 Proof of Lemma 10.7

Denote $\boldsymbol{P}_{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{U}^\top$, for any $\boldsymbol{v} \in \mathbb{R}^n$, we can decompose $\boldsymbol{v}$ into two orthogonal component:

$$\boldsymbol{v} = \boldsymbol{v}_{\boldsymbol{U}} + \boldsymbol{v}_\perp,$$

where $\boldsymbol{v}_{\boldsymbol{U}} = \boldsymbol{P}_{\boldsymbol{U}}\boldsymbol{v}$ and $\boldsymbol{v}_\perp = \boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{v}$. Without loss of generality, assume that $\|\boldsymbol{v}\|_2 = \|\boldsymbol{v}_{\boldsymbol{U}}\|_2^2 + \|\boldsymbol{v}_\perp\|_2^2 = 1$. Thus, we have

$$\begin{aligned}\boldsymbol{v}^\top(\boldsymbol{M}+\boldsymbol{G})\boldsymbol{v} &= (\boldsymbol{v}_{\boldsymbol{U}}+\boldsymbol{v}_\perp)^\top(\boldsymbol{M}+\boldsymbol{G})(\boldsymbol{v}_{\boldsymbol{U}}+\boldsymbol{v}_\perp)\\ &= \boldsymbol{v}_{\boldsymbol{U}}^\top\boldsymbol{M}\boldsymbol{v}_{\boldsymbol{U}} + \boldsymbol{v}_{\boldsymbol{U}}^\top\boldsymbol{G}\boldsymbol{v}_{\boldsymbol{U}} + \boldsymbol{v}_{\boldsymbol{U}}^\top\boldsymbol{G}\boldsymbol{v}_\perp + \boldsymbol{v}_\perp^\top\boldsymbol{G}\boldsymbol{v}_{\boldsymbol{U}} + \boldsymbol{v}_\perp^\top\boldsymbol{G}\boldsymbol{v}_\perp, \end{aligned} \tag{10.33}$$

where the last equation stems from the fact that $\boldsymbol{M} = \mathcal{P}_{\boldsymbol{U}} \boldsymbol{M} \boldsymbol{P}_{\boldsymbol{U}}$ and $\boldsymbol{P}_{\boldsymbol{U}} \boldsymbol{P}_{\boldsymbol{U}_\perp} = \boldsymbol{0}$. Since $\boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{G} \boldsymbol{P}_{\boldsymbol{U}_\perp} = \boldsymbol{0}$, we have

$$\boldsymbol{v}_\perp^\top \boldsymbol{G} \boldsymbol{v}_\perp = \boldsymbol{v}^\top \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{G} \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{v} = 0.$$

Thus, (10.33) is equivalent to

$$\boldsymbol{v}^\top (\boldsymbol{M} + \boldsymbol{G}) \boldsymbol{v} = \boldsymbol{v}_{\boldsymbol{U}}^\top \boldsymbol{M} \boldsymbol{v}_{\boldsymbol{U}} + \boldsymbol{v}_{\boldsymbol{U}}^\top \boldsymbol{G} \boldsymbol{v}_{\boldsymbol{U}} + 2 \boldsymbol{v}_{\boldsymbol{U}}^\top \boldsymbol{G} \boldsymbol{v}_\perp . \tag{10.34}$$

Now let us lower-bound each term on the RHS of (10.34) as follows. First, by the Rayleigh quotient, we have

$$\boldsymbol{v}_{\boldsymbol{U}}^\top \boldsymbol{M} \boldsymbol{v}_{\boldsymbol{U}} \geq \lambda_r \left\| \boldsymbol{v}_{\boldsymbol{U}} \right\|_2^2 , \tag{10.35}$$

and

$$\boldsymbol{v}_{\boldsymbol{U}}^\top \boldsymbol{G} \boldsymbol{v}_{\boldsymbol{U}} \geq \lambda_{\min}(\boldsymbol{G}) \left\| \boldsymbol{v}_{\boldsymbol{U}} \right\|_2^2 \geq - \left\| \boldsymbol{G} \right\|_F \left\| \boldsymbol{v}_{\boldsymbol{U}} \right\|_2^2 . \tag{10.36}$$

Next, by Cauchy-Schwarz inequality,

$$\boldsymbol{v}_{\boldsymbol{U}}^\top \boldsymbol{G} \boldsymbol{v}_\perp \geq - \left\| \boldsymbol{G} \right\|_2 \left\| \boldsymbol{v}_{\boldsymbol{U}} \right\|_2 \left\| \boldsymbol{v}_\perp \right\|_2 \geq - \left\| \boldsymbol{G} \right\|_F \left\| \boldsymbol{v}_{\boldsymbol{U}} \right\|_2 . \tag{10.37}$$

From (10.35), (10.36), and (10.37), we obtain

$$\boldsymbol{v}^\top(\boldsymbol{M}+\boldsymbol{G})\boldsymbol{v} \geq (\lambda_r - \|\boldsymbol{G}\|_F)\,\|\boldsymbol{v}_{\boldsymbol{U}}\|_2^2 - 2\,\|\boldsymbol{G}\|_F\,\|\boldsymbol{v}_{\boldsymbol{U}}\|_2\,. \tag{10.38}$$

Note that $\|\boldsymbol{G}\|_F = \delta$ and the quadratic $g(t) = (\lambda_r - \delta)t^2 - 2\delta t$ is minimized at

$$t_* = \frac{\delta}{\lambda_r - \delta}, \quad g(t_*) = -\frac{\delta^2}{\lambda_r - \delta}.$$

Combining this with (10.38) yields

$$\boldsymbol{v}^\top(\boldsymbol{M}+\boldsymbol{G})\boldsymbol{v} \geq -\frac{2}{\lambda_r}\delta^2,$$

for sufficiently small δ. Let $\boldsymbol{F} = \boldsymbol{G} + \frac{2}{\lambda_r}\delta^2\boldsymbol{I}_n$. Now we can easily verify that $\|\boldsymbol{F}-\boldsymbol{G}\|_F = \mathcal{O}(\delta^2)$ and $\boldsymbol{F} \in \mathcal{F}$.

10.5.7 Proof of Lemma 6.4

We shall show that the matrix $\boldsymbol{E} = \mathcal{P}_r(\boldsymbol{M}+\boldsymbol{F}) - \boldsymbol{M}$ belongs to $\mathcal{E}$ and satisfies

$$\|\boldsymbol{E}-\boldsymbol{F}\|_F = \mathcal{O}(\delta^2). \tag{10.39}$$

First, since $\boldsymbol{F} \in \mathcal{F}_\delta$, $\boldsymbol{M}+\boldsymbol{F}$ must be PSD. Thus, $\mathcal{P}_r(\boldsymbol{M}+\boldsymbol{F})$ is a PSD matrix of rank no greater than r and it admits a rank-r factorization $\mathcal{P}_r(\boldsymbol{M}+\boldsymbol{F}) = \boldsymbol{Z}\boldsymbol{Z}^\top$,

for some $\boldsymbol{Z} \in \mathbb{R}^{n\times r}$. Therefore, by the definition of $\mathcal{E}$,

$$\boldsymbol{E} = \mathcal{P}_r(\boldsymbol{M}+\boldsymbol{F}) - \boldsymbol{M} = \boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{M} \in \mathcal{E}.$$

Next, using (10.17), we have

$$\begin{aligned}\boldsymbol{E} - \boldsymbol{F} &= \mathcal{P}_r(\boldsymbol{M}+\boldsymbol{F}) - \boldsymbol{M} - \boldsymbol{F} \\ &= \boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{F}\boldsymbol{P}_{\boldsymbol{U}_\perp} + \boldsymbol{\mathcal{O}}(\|\boldsymbol{F}\|_F^2).\end{aligned}$$

Since $\boldsymbol{F} \in \mathcal{F}_\delta$ implies $\boldsymbol{P}_{\boldsymbol{U}_\perp}\boldsymbol{F}\boldsymbol{P}_{\boldsymbol{U}_\perp} = \boldsymbol{\mathcal{O}}(\|\boldsymbol{F}\|_F^2)$, we conclude that $\boldsymbol{E} - \boldsymbol{F} = \boldsymbol{\mathcal{O}}(\|\boldsymbol{F}\|_F^2)$.

Chapter 11: Adaptive Step Size Momentum Method For Deconvolution[1]

In this chapter, we introduce an adaptive step size schedule that can significantly improve the convergence rate of momentum method for deconvolution applications. We provide analysis to show that the proposed method can asymptotically recover the optimal rate of convergence for first-order gradient methods applied to minimize smooth convex functions. In a convolution setting, we demonstrate that our adaptive scheme can be implemented efficiently without adding computational complexity to traditional gradient schemes.

11.1 Introduction

Deconvolution is the process of reversing the effects of convolution [227]. It is widely used in the areas of signal processing and image processing [10,19]. In image processing, this term also refers to recovering the original image by deblurring [8]. Recently there has been an increasing interest in machine learning approaches for deconvolution including nonnegative matrix factorization [155], sparse coding [122, 185], convolutional dictionary learning [39, 233].

[1]This work has been published as: "Adaptive Step Size Momentum Method For Deconvolution", In Proceedings of IEEE Statistical Signal Processing Workshop (SSP), pp. 438-442. IEEE, 2018.

Deconvolution is usually performed by representing the convolution in the form of a linear shift-invariant operator and utilize a minimum mean square error as an optimization criterion. From machine learning perspective, the objective function can also be extended to other loss functions like Hinge loss or logistic regression cost function. Deconvolution can be done on either the time domain using circulant matrices or the frequency domain by computing the Fourier Transform [99]. A major challenge of this inverse problem is the ill-posed nature of continuous data that results in ill-conditioned matrices in the optimization [157]. Several techniques have been proposed to accomplish this using regularization theory. One direct way is to compute the closed-form solution of the problem. However, this approach is often inefficient due to the computational complexity of the inverse operator, and more importantly, it only works for apparently simple objective functions [8]. A more common method is to use iterative algorithms, in which various optimization techniques can be exploited to find a close approximation of the solution. With the increasing number of large-scale problems, this approach have been shown to be very well suited to deconvolution. Besides, other deconvolution techniques include recursive filtering [157], wavelets [65], and neural networks [229].

The most widely used among iterative algorithms for deconvolution is gradient descent. Although this method suffers from the slow convergence rate of first-order methods, its low cost and simplicity turn out to be very useful in practice. On the other hand, second-order methods such as Newton-Raphson obtain a rapid convergence rate but require the computation of the Hessian and its inverse, which can be prohibitively expensive for large scale problems [24]. To compromise, momen-

tum method has been proposed to accelerate the convergence of gradient descent while remaining the computational complexity. This slight modification of gradient descent was shown to achieve a fast convergence rate on minimizing a smooth convex function [159]. Nevertheless, while multiple approaches are available for choosing optimal step sizes in gradient descent (e.g., backtracking line search), little is known for step size selection in momentum method when prior knowledge of the function curvature is limited.

To address this issue, we propose an adaptive schedule that uses the gradient information to compute the step size for momentum method at each iteration accordingly. In a convolution setting, the special structure of the objective function allows us to implement the algorithm efficiently without heavy computations of the Hessian. We provide analysis to show that our method asymptotically recovers the optimal convergence rate determined by the Hessian at the solution. Compared to gradient descent methods, the proposed method requires only twice as many as the number of operations per iteration, while dramatically accelerates the convergence in many cases when the objective function is ill-conditioned in general but locally well-conditioned at the solution. Lastly, we present a numerical evaluation that verifies the effectiveness of the proposed approach and suggest potential applications to other domains.

11.2 Preliminary

Consider the problem of minimizing a twice differentiable, smooth and strongly convex function $f(x) : \mathbb{R}^d \to \mathbb{R}$. In particular, $lI \preceq \nabla^2 f(x) \preceq LI$, $\forall x$. We shall denote by x^* the unique solution of this optimization problem and $f^* = f(x^*)$. We further assume that λ_1 and λ_d are the largest and smallest eigenvalues of the Hessian at x^*, respectively. Thus, we define the global condition number of f as $r = \frac{L}{l}$, and the local condition number of $\nabla^2 f^*$ as $\kappa = \frac{\lambda_1}{\lambda_d}$. For quadratics, these two number are the same. However, for non-quadratic functions, κ is smaller than r. Exploiting the gap between k and r is often an efficient way to accelerate the convergence of iterative methods.

In gradient descent method, the solution is initialized to $x = x^{(0)}$ and the following step is used to update x:

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)}). \tag{11.1}$$

Various algorithms have been proposed to choose the step size $\alpha^{(k)}$ in order to obtain an optimal convergence rate. One notable result from Nesterov regarding fixed step size gradient descent methods is given by Theorem 11.1.

Theorem 11.1. *(Theorem 2.1.15 in [160]). The gradient descent method with fixed step size* $\alpha^{(k)} = \frac{2}{L+l}$ *has a global linear convergence rate of* $R = \frac{r-1}{r+1} = \frac{L-l}{L+l}$,

i.e.,

$$f(x^{(k+1)}) - f^* \leq \frac{L}{2}\left(\frac{L-l}{L+l}\right)^{2k} \left\|x^{(0)} - x^*\right\|^2.$$

Alternatively, adaptive schedules like exact and inexact line search are generally preferred in practice. It has recently been shown to converge at the same rate R for smooth convex functions on the worst-case scenario [54]. However, beyond the worst-case scenario, we should note that the asymptotic convergence rate is generally better for non-quadratic functions, where the objective function is locally well-conditioned, and the asymptotic convergence rate is defined by the local condition number of the Hessian at the solution: $K = \frac{\kappa-1}{\kappa+1}$, which is smaller than R.

Momentum method adds a second term from the previous iterate to the update equation of gradient descent

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)}\nabla f(x^{(k)}) + \beta^{(k)}(x^{(k)} - x^{(k-1)}). \tag{11.2}$$

In [166], Polyak showed that this method achieves a faster convergence rate of $\frac{\sqrt{r}-1}{\sqrt{r}+1}$ on a quadratic by setting

$$\alpha^{(k)} = \left(\frac{2}{\sqrt{L}+\sqrt{l}}\right)^2,\ \beta^{(k)} = \left(\frac{\sqrt{L}-\sqrt{l}}{\sqrt{L}+\sqrt{l}}\right)^2. \tag{11.3}$$

It is noteworthy that analyses of convergence in this case usually involve performing a change of basis on the domain size $y^{(k)} = U^\top(x^{(k)} - x^*)$, where U comes from the

eigenvalue decomposition $\nabla^2 f^* = U\Lambda U^\top$. The convergence rate is then defined by the slowest decreasing component in $y^{(k)}$, denoted by $y_j^{(k)}$. Similar to gradient descent, fixing the momentum step size does not recover the optimal convergence rate for non-quadratic smooth convex objective function. An in-depth analysis on different momentum regimes which is based on the behavior of second-order dynamic systems is discussed in [164]. The authors also suggested a restart strategy in order to achieve an even faster convergence rate that depends on the condition number of Hessian at solution, $\tau = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. However, the proposed algorithms are based on Nesterov's Accelerated Gradient method, a variant of momentum method, and hence its motivation is rather the restarting mechanism than choosing the optimal step sizes.

11.3 Problem Formulation

In deconvolution, we are given a training set of $\{\boldsymbol{x}_m, \boldsymbol{y}_m\}_{m=1}^M$ and the objective is to learn a convolution kernel $\boldsymbol{w}$ to minimize a cost function $f(\boldsymbol{w}) = \sum_{m=1}^M \mathcal{C}(\boldsymbol{x}_m * \boldsymbol{w}, \boldsymbol{y}_m) + \Omega(\boldsymbol{w})$, where Ω is a regularization term. Assume all training examples are of the same dimension n and are zero-padded at both ends, we can denote $\boldsymbol{x}_m = [x_m(1), \ldots, x_m(n)]^\top$, $\boldsymbol{y}_m = [\boldsymbol{y}_m(1), \ldots, \boldsymbol{y}_m(n)]^\top$, and $\boldsymbol{w} = [w(1), \ldots, w(h)]^\top$, where h is the window size. Let $\boldsymbol{x}_{mt} = [x_m(t), x_m(t-1), \ldots, x_m(t-h+1)]^\top$ be the tth sliding window segment in the mth signal. If $\mathcal{C}$ can be broken down to each

sliding window, the objective function is rewritten as

$$f(\boldsymbol{w}) = \sum_{m=1}^{M} \sum_{t=1}^{n} c(\boldsymbol{w}^\top \boldsymbol{x}_{mt}, \boldsymbol{y}_m(t)) + \Omega(\boldsymbol{w}). \tag{11.4}$$

We make a further assumption that the cost function $c(\boldsymbol{a}, \boldsymbol{b})$ is smooth convex and the regularizer $\Omega(\boldsymbol{w})$ is smooth and strongly convex, and both are twice differentiable with respect to $\boldsymbol{a}$ and $\boldsymbol{w}$, i.e., $0 \preceq \frac{\partial^2 c}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\top} \preceq \mu I$ and $\lambda I \preceq \frac{d^2 \Omega}{d\boldsymbol{w} d\boldsymbol{w}^\top} \preceq \gamma I$, for $\mu, \lambda, \gamma > 0$. Note that $c(\boldsymbol{a}, \boldsymbol{b})$ can be a distance metric (e.g., $\|\boldsymbol{a} - \boldsymbol{b}\|^2$), a divergence metric (e.g., $\sum_i \left(b_i \log \frac{b_i}{a_i} + a_i - b_i \right)$ where $a_i, b_i > 0$) or a more general loss (e.g., $\log(\sum_i e^{a_i}) - \sum_i a_i b_i$ where $b_i \in \{0, 1\}$). Consider a logistic loss with L2-regularization for example, $c(a, b) = \log(1 + e^a) - ab$, where $b \in \{0, 1\}$, and $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|^2$, the aforementioned assumption is supported by $0 \leq \frac{\partial^2 c}{\partial a^2} = \frac{\partial c}{\partial a}(1 - \frac{\partial c}{\partial a}) \leq \frac{1}{4}$ and $\frac{d^2 \Omega}{d\boldsymbol{w} d\boldsymbol{w}^\top} = \lambda I$. For simplicity, we provide analysis for the case where the cost function parameters $\boldsymbol{a}, \boldsymbol{b}$ are scalars. From (11.4), we obtain

$$\nabla^2 f(\boldsymbol{w}) = \sum_{m=1}^{M} \sum_{t=1}^{n} \frac{\partial^2 c}{\partial (\boldsymbol{w}^\top \boldsymbol{x}_{mt})^2} \boldsymbol{x}_{mt} \boldsymbol{x}_{mt}^\top + \frac{d^2 \Omega}{d\boldsymbol{w} d\boldsymbol{w}^\top}. \tag{11.5}$$

In this setting, the special structure of the Hessian recalls the autocorrelation $R_{\hat{x}_m} = X_m^\top X_m$, where the circulant matrix $X_m = [\boldsymbol{x}_{m1}^\top, \ldots, \boldsymbol{x}_{m(n+h-1)}^\top]$ is obtained from padding zero to $\boldsymbol{x}_m$ and $\hat{\boldsymbol{x}}_m$ is the time-reversed version of $\boldsymbol{x}_m$. If all signals

are normalized to zero mean and unit variance, the Hessian can be bounded by

$$\lambda I \preceq \nabla^2 f(\boldsymbol{w}) \preceq \mu \sum_{m=1}^{M} R_{\hat{x}_m} + \gamma I \quad \forall \boldsymbol{w}. \tag{11.6}$$

Since the power spectrum of $\hat{x}_m$ can be expressed as the Fourier Transform of its autocorrelation function, the maximum eigenvalue of $R_{\hat{x}_m}$ is also the maximum power spectrum $S_{\hat{x}_m}(0)$. Thus, (11.6) provides us with a decent estimate of the function parameters: $l = \lambda$ and $L = \mu \sum_{m=1}^{M} S_{\hat{x}_m}(0) + \lambda$. In this estimation, the lower bound depends on the choice of the regularization factor, while the upper bound depends on the data itself.

11.4 Adaptive Step Size Scheme

Motivated by line search approach in gradient descent, this section presents an adaptive schedule to choose the optimal value of step size for each momentum iteration. First, we notice that the optimal convergence rate for momentum, and in general for first-order methods, depends on the condition number of the Hessian at the solution. Indeed, asymptotic analyses of convergence often assume the function can be locally approximated by a quadratic in the region near the optimum, and consider the rate of convergence inside this region. In case of gradient descent, recovering the optimal rate is practically done by backtracking line search. Another inexact line search method stems from second-order Taylor expansion of the objective function $f(\boldsymbol{w} - \alpha \nabla f(\boldsymbol{w})) \approx f(\boldsymbol{w}) - \alpha \nabla f(\boldsymbol{w})^\top \nabla f(\boldsymbol{w}) + \frac{1}{2}\alpha^2 \nabla f(\boldsymbol{w})^\top \nabla^2 f(\boldsymbol{w}) \nabla f(\boldsymbol{w})$.

The superscripts are omitted for brevity. The step size at each iteration is then determined by minimizing this quadratic function with respect to α, yielding

$$\alpha = \frac{\nabla f(\boldsymbol{w})^\top \nabla f(\boldsymbol{w})}{\nabla f(\boldsymbol{w})^\top \nabla^2 f(\boldsymbol{w}) \nabla f(\boldsymbol{w})}. \tag{11.7}$$

Since quadratic functions have the same Hessian everywhere, it follows from Section 11.2 that the resulting iterates obtain the optimal asymptotic convergence rate K inside the quadratic region near the solution.

Naturally, we bring this intuition to the updates in momentum method. Let $\Delta \boldsymbol{w}^{(k)} = \boldsymbol{w}^{(k)} - \boldsymbol{w}^{(k-1)}$, $D_k = [\nabla f(\boldsymbol{w}^{(k)}), -\Delta \boldsymbol{w}^{(k)}]$, and $\boldsymbol{\eta}^{(k)} = [\alpha^{(k)}, \beta^{(k)}]^\top$. From (11.2), we can approximate

$$f(\boldsymbol{w}^{(k)} - D_k \boldsymbol{\eta}^{(k)}) \approx f(\boldsymbol{w}^k) - \nabla f(\boldsymbol{w}^{(k)})^\top D_k \boldsymbol{\eta}^{(k)} + \frac{1}{2} {\boldsymbol{\eta}^{(k)}}^\top D_k^\top \nabla^2 f(\boldsymbol{w}^{(k)}) D_k \boldsymbol{\eta}^{(k)}.$$

Minimizing this quadratic function with respect to $\boldsymbol{\eta}^{(k)}$ yields

$$\boldsymbol{\eta}^{(k)} = \left(D_k^\top \nabla^2 f(\boldsymbol{w}^{(k)}) D_k \right)^{-1} D_k^\top \nabla f(\boldsymbol{w}^{(k)}). \tag{11.8}$$

We can further simplify (11.8) as follows

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \nabla f^\top \nabla^2 f \nabla f & -\Delta \boldsymbol{w}^\top \nabla^2 f \nabla f \\ -\Delta \boldsymbol{w}^\top \nabla^2 f \nabla f & \Delta \boldsymbol{w}^\top \nabla^2 f \Delta \boldsymbol{w} \end{bmatrix}^{-1} \begin{bmatrix} \nabla f^\top \nabla f \\ -\Delta \boldsymbol{w}^\top \nabla f \end{bmatrix}$$

Note that the inversion in this equation only involves a 2×2 matrix, and should not

Algorithm 11.1 Adaptive step size scheme for momentum.

Given initial guess $\boldsymbol{w}^{(0)}$ and $\boldsymbol{w}^{(1)}$.
repeat for $k = 1, 2, \ldots$
  $\Delta \boldsymbol{w} = \boldsymbol{w}^{(k)} - \boldsymbol{w}^{(k-1)}$ ▷ $O(h)$
  $\nabla f = \sum_{m,t} \frac{\partial c}{\partial (\boldsymbol{w}^\top \boldsymbol{x}_{mt})} \boldsymbol{x}_{mt} + \lambda \boldsymbol{w}$ ▷ $O(Mnh)$
  for $m = 1, \ldots, M$, $t = 1, \ldots, n$ **do** ▷ $O(Mnh)$
    $p_{mt} = \boldsymbol{x}_{mt}^\top \nabla f$, $q_{mt} = \boldsymbol{x}_{mt}^\top \Delta \boldsymbol{w}$
    $c_{mt} = \partial^2 c / \partial (\boldsymbol{w}^\top \boldsymbol{x}_{mt})^2$
  $u = \nabla f^\top \nabla f$, $v = \Delta \boldsymbol{w}^\top \nabla f$, $t = \Delta \boldsymbol{w}^\top \Delta \boldsymbol{w}$ ▷ $O(h)$
  $a = \sum_{m=1}^{M} \sum_{t=1}^{n} c_{mt} p_{mt}^2 + \lambda u$ ▷ $O(Mn)$
  $b = \sum_{m=1}^{M} \sum_{t=1}^{n} c_{mt} p_{mt} q_{mt} + \lambda v$ ▷ $O(Mn)$
  $d = \sum_{m=1}^{M} \sum_{t=1}^{n} c_{mt} q_{mt}^2 + \lambda t$ ▷ $O(Mn)$
  $\alpha^{(k)} = \frac{du - bv}{ad - b^2}$, $\beta^{(k)} = \frac{bu - av}{ad - b^2}$ ▷ $O(1)$
  Update $\boldsymbol{w}^{(k+1)}$ using (11.2). ▷ $O(h)$
until convergence

be confused with the inversion in Newton-Raphson method. More interestingly, computing this matrix only requires the same complexity as computing the gradient thanks to the decomposition of $\nabla^2 f$ into multiple terms of the form $\boldsymbol{x}\boldsymbol{x}^\top$. We propose the adaptive step size momentum method in Algorithm 11.1, with L2-regularization for simplicity.

The resulting iterates obtain a provably asymptotic convergence rate $\tau = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}-1}$. The detailed proof is not given here due to space limitation. The intuition is each iteration Algorithm 11.1 decreases the objective function more than that of fixed step size momentum chosen by (11.3). Therefore, inside the optimal region, it converges at least as fast as fixed step size chosen by the local parameters. Although the behavior outside the optimal region is unclear, this adaptive scheme is often helpful in practice.

Table 11.1: Computational complexity of fixed step size gradient (GD), adaptive step size gradient (AGD), fixed step size momentum (MO), adaptive step size momentum (AMO), and Newton's method. ϵ is the relative accuracy.

Method	# Ops. / Iter.	Cvg. rate	# Iters. needed
GD	$O(Mnh)$	$\frac{r-1}{r+1}$	$\frac{r+1}{2}\log(1/\epsilon)$
AGD	$O(Mnh)$	$\frac{\kappa-1}{\kappa+1}$	$\frac{\kappa+1}{2}\log(1/\epsilon)$
MO	$O(Mnh)$	$\frac{\sqrt{r}-1}{\sqrt{r}+1}$	$\frac{\sqrt{r}+1}{2}\log(1/\epsilon)$
AMO	$O(Mnh)$	$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$	$\frac{\sqrt{\kappa}+1}{2}\log(1/\epsilon)$
Newton	$O(Mnh^3)$	quadratic	$O(1)$

To simplify the complexity analysis, we assume the calculation of derivatives of c and Ω is $O(1)$. Table 11.1 shows the computational complexity per iteration of the proposed approach is the same as other methods, but it requires the least number of iterations to reach a certain accuracy to the solution. The computation can even be more efficient if the window size is large enough ($h \approx n$), by using the Fast Fourier Transform to obtain $O(Mn \log n)$ complexity per iteration.

11.5 Numerical Example

In this experiment, we consider a convolutive logistic regression model for recognizing a sequence of handwritten digits. Our goal is to illustrate the convergence of the proposed algorithm and compare it with the theoretical analysis in the previous section.

Setting. From MNIST database, we generate a dataset of $M = 10$ composite images as follows. Each image of size 28×150 is created by sequentially adding four

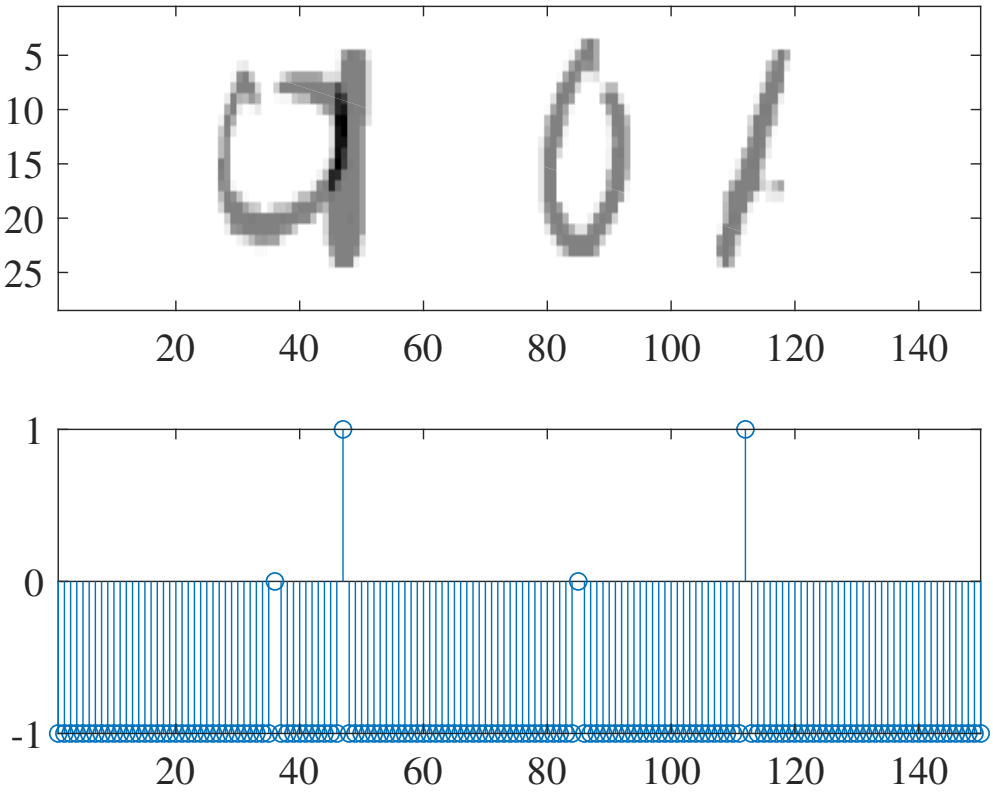


Figure 11.1: Top - an image generated by randomly inserting a sequence of $0, 1, 0, 1$. Bottom - the corresponding label series.

28×28 digit images to a zero background such that the bottom left corner of the ith digit is chosen uniformly between positions $28(i-1)+1$ and $28i$ along the width of the composite image (see Fig. 11.1). We create the feature vector $\boldsymbol{x}_{mt}$ of $h = 784$ elements by vectorizing the 28×28 window centered at tth position along the width of the mth image ($n = 150$). For the purpose of illustration, we only consider images with two digits 0 and 1. Thus, we aim to learn a classifier $\boldsymbol{w}$ for $C = 3$ classes 0, 1 and -1 (for non-digit positions). The parameter $\boldsymbol{w}$ thereupon has 2352 elements ($= 3 \times 28 \times 28$), making the Hessian exceedingly large and infeasible to apply Newton's method. Finally, the label $\boldsymbol{y}_m(t)$ is determined by checking whether the window centered at tth position matches exactly with the digit positions. We represent the label as a vector of class membership $\boldsymbol{y}_m(t) = [y_{mt1}, \ldots, y_{mtC}]^\top$.

Recall the multinomial logit-model is given by

$$p_{mtc} = P(y_{mtc}|\boldsymbol{x}_{mt}, \boldsymbol{w}) = e^{y_{mtc}\boldsymbol{w}_c^\top \boldsymbol{x}_{mt}} / (\sum_{j=1}^{C} e^{\boldsymbol{w}_j^\top \boldsymbol{x}_{mt}})$$

and the cost $c(\boldsymbol{a}, \boldsymbol{b}) = \log(\sum_{c=1}^{C} e^{a_c}) - \sum_{c=1}^{C} a_c b_c$. The Hessian can be extended from (11.5) as

$$\nabla^2 f = \frac{1}{Mn} \sum_{m,n} (\Lambda_{\boldsymbol{p}_{mt}} - \boldsymbol{p}_{mt}\boldsymbol{p}_{mt}^\top) \otimes \boldsymbol{x}_{mt}\boldsymbol{x}_{mt}^\top,$$

where $\boldsymbol{p}_{mt} = [p_{mt1}, \ldots, p_{mtC}]^\top$ (see [22]). Since we have $\Lambda_{\boldsymbol{p}_{mt}} - \boldsymbol{p}_{mt}\boldsymbol{p}_{mt}^\top \preceq \frac{1}{2}(I_C - \frac{\mathbf{1}\mathbf{1}^\top}{C+1})$, a rough estimate of the Lipschitz constant is $L = \frac{1}{2Mn} \sum_m S_{\hat{x}_m}(0)$. The strong convexity constant l is controlled by the L2-regularization factor $\lambda = 10^{-2}$. Thereupon, we implement four other methods in comparison with our proposed

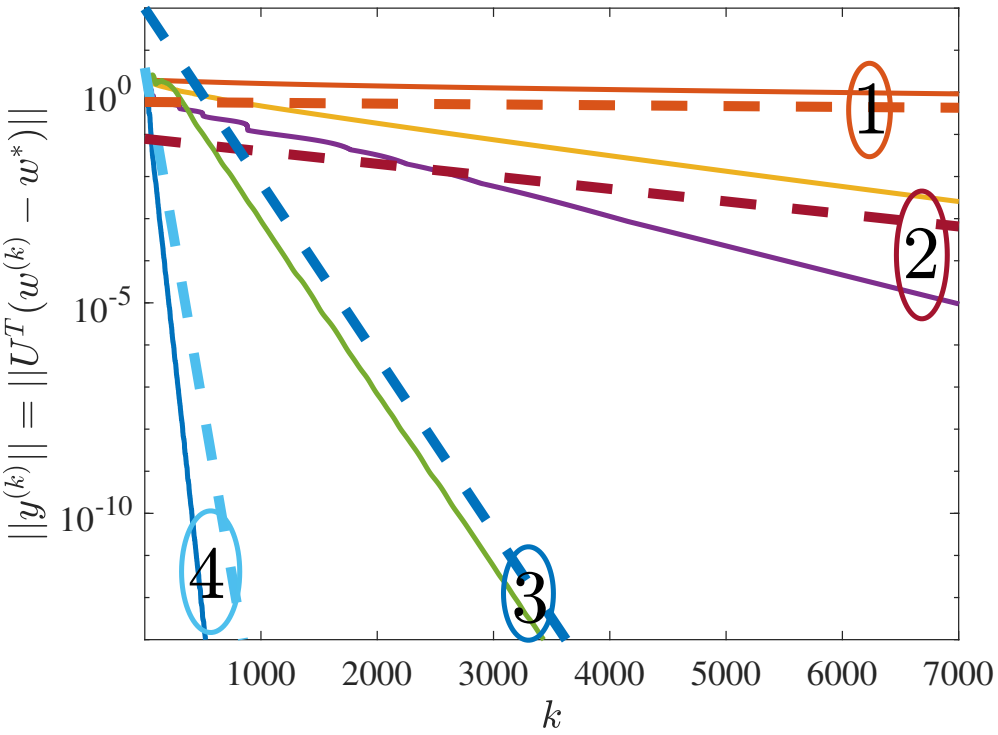


Figure 11.2: The log-scale decrease in the distance to the solution on domain value side through iterations.

method, namely fixed step size gradient descent, adaptive step size gradient descent, fixed step size momentum, and gradient descent with backtracking line search. For fixed step size schemes, we use the optimal step sizes described in Theorem 11.1 and Equation (11.3). For backtracking line search, we set the parameters $\alpha = 0.2, \beta = 0.5$. Our adaptive step size schedules require no tuning parameters.

Results and analysis. Figure 11.2 compares the empirical convergence of the five methods in terms of the distance to the solution. The dash lines are added to the plot in order to depict the theoretical convergence rate corresponding

to the global and local condition numbers given in Table 11.1. Not surprisingly all the methods match their theoretical convergence rates. The convergence of adaptive step size momentum deems to be slightly faster than the optimal rate at the solution (group 4), and clearly outperforms the other four methods. Adaptive schemes applied to gradient descent also results in a competitive convergence to backtracking line search, i.e., purple line versus yellow line (group 2). Fixed step size gradient descent and momentum method converge almost at the rate predicted by the analysis (group 1 and 3). Obviously, those approaches are slower than their adaptive versions because they only depend on the global parameters of the objective function and cannot recover the optimal rate at the solution. For quadratic objectives, this distinction is occluded by the fact that the Hessian is constant everywhere.

11.6 Conclusion

To conclude, we proposed an adaptive schedule for choosing step size in momentum method, under deconvolution settings. It can be readily implemented without adding computational complexity to fixed step size schemes. We showed that our method outperforms other aforementioned iterative methods in terms of convergence rate. It is promising that the proposed approach can be applied to a wide range of problems in the domain of digital signal and image processing.

11.7 Appendix

In this section, we consider a simple quadratic function

$$f(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{d} \lambda_i x_i^2 = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Lambda} \boldsymbol{x}, \tag{11.9}$$

$$\nabla f(\boldsymbol{x}) = \boldsymbol{\Lambda}\boldsymbol{x}, \tag{11.10}$$

$$\nabla^2 f(\boldsymbol{x}) = \boldsymbol{\Lambda}. \tag{11.11}$$

The results can be generalized to asymptotic analysis of other convex functions based on the following proposition

Proposition 11.1. *Let (f_k) be the sequence defined by the recursion $f_{k+1} = af_k + bf_k^2$, for $k = 1, 2, \ldots$. If $a < 1$ and $f_1 < \frac{1-a}{b}$, then (f_k) converges to 0 at asymptotic rate a.*

Proof. Since (f_k) is strictly decreasing, it is easy to show that , with $a < 1$,

- If $f_1 > \frac{1-a}{b}$, (f_k) diverges.
- If $f_1 = \frac{1-a}{b}$, $(f_k) = \frac{1-a}{b}$.
- If $f_1 < \frac{1-a}{b}$, (f_k) converges to 0.

Consider the case when (f_k) converges to 0. There must exist k_0 such that $f_k < \frac{a(1-a)}{b}$, for all $k \geq k_0$.

Suppose that $f_1 = \alpha\frac{1-a}{b}$, where $0 < \alpha < 1$. Let us define a sequence (h_k) as

$h_k = \frac{1}{f_1 a^{k-1}} f_k$. Then for $k \geq k_0$

$$h_k = \frac{1}{f_1 a^{k-1}} f_k < \frac{1}{f_1 a^{k-1}} \frac{a(1-a)}{b} = \frac{1}{\alpha a^{k-2}}. \tag{11.12}$$

The recursion for (h_k) is given by

$$\begin{cases} h_1 = 1, \\ h_{k+1} = h_k + \alpha(1-a)a^{k-2}h_k^2. \end{cases}$$

Notice that (h_k) is also strictly increasing, and the following inequalities hold

$$\begin{aligned}
\Rightarrow \quad & \alpha(1-a)a^{k-2} = \frac{h_{k+1}-h_k}{h_k^2} > \frac{h_{k+1}-h_k}{h_{k+1}h_k} = \frac{1}{h_k} - \frac{1}{h_{k+1}} \\
\Rightarrow \quad & \sum_{i=k_0}^{k-1} \alpha(1-a)a^{i-2} > \sum_{i=k_0}^{k-1} \left(\frac{1}{h_i} - \frac{1}{h_{i+1}} \right) \\
\Rightarrow \quad & \alpha(1-a)a^{k_0-2} \sum_{j=0}^{k-1-k_0} a^j > \frac{1}{h_{k_0}} - \frac{1}{h_k} \\
\Rightarrow \quad & \alpha(1-a)a^{k_0-2} \frac{1-a^{k-k_0}}{1-a} > \frac{1}{h_{k_0}} - \frac{1}{h_k} \\
\Rightarrow \quad & \frac{1}{h_k} > \frac{1}{h_{k_0}} - \alpha a^{k_0-2}(1-a^{k-k_0}) \\
\Rightarrow \quad & h_k < \frac{1}{\left(\frac{1}{h_{k_0}} - \alpha a^{k_0-2}\right) + \alpha a^{k-2}}.
\end{aligned}$$

From (11.12), the sequence defined by the RHS must converge to a constant $\frac{1}{\frac{1}{h_{k_0}} - \alpha a^{k_0-2}}$. Consequently, (h_k) is upper-bounded by this sequence and also converges. Finally, we obtain $\lim h_k = \lim \frac{a}{f_1} \frac{f_k}{a^k} < \infty$, yielding the asymptotic conver-

gence rate of (f_k) to 0 is a. □

11.7.1 Fixed Step Size Gradient Descent

From the update $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{x}^{(k)} - \alpha \boldsymbol{\Lambda} \boldsymbol{x}^{(k)}$, we have

$$\begin{aligned} f(\boldsymbol{x}^{(k+1)}) &= \frac{1}{2}(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{\Lambda} \boldsymbol{x}^{(k)})^\top \boldsymbol{\Lambda} (\boldsymbol{x}^{(k)} - \alpha \boldsymbol{\Lambda} \boldsymbol{x}^{(k)}) \\ &= \frac{1}{2}(\boldsymbol{x}^{(k)})^\top (\boldsymbol{I} - \alpha \boldsymbol{\Lambda}) \boldsymbol{\Lambda} (\boldsymbol{I} - \alpha \boldsymbol{\Lambda}) \boldsymbol{x}^{(k)} \\ &= \frac{1}{2}(\boldsymbol{\Lambda}^{1/2} \boldsymbol{x}^{(k)})^\top (I - \alpha \boldsymbol{\Lambda})^2 (\boldsymbol{\Lambda}^{1/2} \boldsymbol{x}^{(k)}) \\ &\leq \frac{1}{2} \left\| \boldsymbol{I} - \alpha \boldsymbol{\Lambda} \right\|^2 . \left\| \boldsymbol{\Lambda}^{1/2} \boldsymbol{x}^{(k)} \right\|^2 = \max_i (1 - \alpha \lambda_i)^2 . f(\boldsymbol{x}^{(k)}). \end{aligned}$$

By setting $\alpha = \frac{2}{\lambda_1 + \lambda_d}$, we obtain

$$f(\boldsymbol{x}^{(k+1)}) \leq \left(\frac{\lambda_1 - \lambda_d}{\lambda_1 + \lambda_d} \right)^2 f(\boldsymbol{x}^{(k)}).$$

11.7.2 Fixed Step Size Momentum Method

From the update $\boldsymbol{x}^{(k+1)} = x^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)}) + \beta (\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)})$, we have

$$\boldsymbol{y}^{(k+1)} = \begin{bmatrix} \boldsymbol{x}^{(k+1)} \\ \boldsymbol{x}^{(k)} \end{bmatrix} = \begin{bmatrix} (1+\beta)\boldsymbol{I} - \alpha \boldsymbol{\Lambda} & -\beta \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^{(k)} \\ \boldsymbol{x}^{(k-1)} \end{bmatrix} = \boldsymbol{M} \boldsymbol{y}^{(k)}$$

and

$$
\begin{aligned}
f_{k+1} = f(\boldsymbol{x}^{(k+1)}) + f(\boldsymbol{x}^{(k)}) &= \frac{1}{2}\boldsymbol{y}^{(k+1)^\top}\begin{bmatrix}\boldsymbol{\Lambda} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}\end{bmatrix}\boldsymbol{y}^{(k+1)}\\
&= \frac{1}{2}\boldsymbol{y}^{(k)^\top}\boldsymbol{M}^\top\hat{\boldsymbol{\Lambda}}\boldsymbol{M}\boldsymbol{y}^{(k)} = \dots\\
&= \frac{1}{2}\boldsymbol{y}^{(1)^\top}\boldsymbol{M}^{k^\top}\hat{\boldsymbol{\Lambda}}\boldsymbol{M}^k\boldsymbol{y}^{(1)}\\
&= \frac{1}{2}(\hat{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{y}^{(1)})^\top\Big(\hat{\boldsymbol{\Lambda}}^{-1/2}\boldsymbol{M}^{k^\top}\hat{\boldsymbol{\Lambda}}\boldsymbol{M}^k\hat{\boldsymbol{\Lambda}}^{-1/2}\Big)(\hat{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{y}^{(1)})\\
&= \frac{1}{2}(\hat{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{y}^{(1)})^\top\Big((\hat{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{M}^k\hat{\boldsymbol{\Lambda}}^{-1/2})^\top(\hat{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{M}^k\hat{\boldsymbol{\Lambda}}^{-1/2})\Big)(\hat{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{y}^{(1)})\\
&\leq \frac{1}{2}\left\|\hat{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{M}^k\hat{\boldsymbol{\Lambda}}^{-1/2}\right\|^2\left\|\hat{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{y}^{(1)}\right\|^2 = \left\|\hat{\boldsymbol{\Lambda}}^{1/2}M^k\hat{\boldsymbol{\Lambda}}^{-1/2}\right\|^2 f_1\\
&\leq \Big(\left\|\hat{\boldsymbol{\Lambda}}^{1/2}\right\|\left\|\boldsymbol{M}^k\right\|\left\|\hat{\boldsymbol{\Lambda}}^{-1/2}\right\|\Big) f_1 = \left\|\boldsymbol{M}^k\right\|^2\frac{\lambda_1^2}{\lambda_d^2}f_1.
\end{aligned}
$$

Since $\lim_{k\to\infty}\frac{\left\|\boldsymbol{M}^k\right\|^2}{\rho(\boldsymbol{M})^k} = 1$,[2] the spectral radius $\rho(\boldsymbol{M}) = \max_j\{|\lambda_j(\boldsymbol{M})|\}$ determines the convergence rate of the series (f_k). Recall that $\boldsymbol{M} = \begin{bmatrix}(1+\beta)I - \alpha\boldsymbol{\Lambda} & -\beta\boldsymbol{I}\\ \boldsymbol{I} & \boldsymbol{0}\end{bmatrix}$. We define the permutation π such that

$$
\pi(j) = \begin{cases} 2j-1 \text{ if } j \leq d,\\ 2j - 2d \text{ otherwise.}\end{cases}
$$

[2]Gelfand's formula.

Then

$$\boldsymbol{M} \sim \boldsymbol{P}_\pi \boldsymbol{M} \boldsymbol{P}_\pi^\top = \begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_2 & \dots & \boldsymbol{0} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{M}_d \end{bmatrix}$$

is a block diagonal matrix with eigenvalues are simply those of $\boldsymbol{M}_1, \boldsymbol{M}_2, \dots, \boldsymbol{M}_d$. For any $j = 1, \dots, d$, the eigenvalues of $\boldsymbol{M}_j$ are the root of the characteristic polynomial $\sigma^2 - (1 + \beta - \alpha\lambda_j)\sigma + \beta$. Since $\alpha = \left(\frac{2}{\sqrt{\lambda_1}+\sqrt{\lambda_d}}\right)^2, \beta = \left(\frac{\sqrt{\lambda_1}-\sqrt{\lambda_d}}{\sqrt{\lambda_1}+\sqrt{\lambda_d}}\right)^2$, the two complex roots are given by

$$\sigma_{j_1,j_2} = \frac{1}{2}\left(1 + \beta - \alpha\lambda_j \pm \sqrt{(1 + \beta - \alpha\lambda_j)^2 - 4\beta}\right).$$

It follows that the magnitudes of all eigenvalues are equal to $\sqrt{\beta}$. Thus $\rho(\boldsymbol{M}) = \sqrt{\beta}$.

11.7.3 Adaptive Step Size Gradient Descent

From (11.9), we have

$$f(\boldsymbol{x}^{(k+1)}) = f(\boldsymbol{x}^{(k)}) - \alpha_k \nabla f(\boldsymbol{x}^{(k)})^\top \nabla f(\boldsymbol{x}^{(k)}) + \frac{1}{2}\alpha_k^2 \nabla f(\boldsymbol{x}^{(k)})^\top \nabla^2 f(\boldsymbol{x}^{(k)}) \nabla f(\boldsymbol{x}^{(k)}).$$

Substituting $\alpha_k = \frac{\nabla f(\boldsymbol{x}^{(k)})^\top \nabla f(\boldsymbol{x}^{(k)})}{\nabla f(\boldsymbol{x}^{(k)})^\top \nabla^2 f(\boldsymbol{x}^{(k)}) \nabla f(\boldsymbol{x}^{(k)})}$, we obtain

$$\begin{aligned}
f(\boldsymbol{x}^{(k+1)}) &= f(\boldsymbol{x}^{(k)}) - \frac{1}{2} \frac{\left(\nabla f(\boldsymbol{x}^{(k)})^\top \nabla f(\boldsymbol{x}^{(k)}) \right)^2}{\nabla f(\boldsymbol{x}^{(k)})^\top \nabla^2 f(\boldsymbol{x}^{(k)}) \nabla f(\boldsymbol{x}^{(k)})} \\
&= f(\boldsymbol{x}^{(k)}) - \frac{1}{2} \frac{\left(\boldsymbol{x}^{(k)\top} \boldsymbol{\Lambda}^2 \boldsymbol{x}^{(k)} \right)^2}{\boldsymbol{x}^{(k)\top} \boldsymbol{\Lambda}^3 \boldsymbol{x}^{(k)}} \\
&= \left(1 - \frac{\left(\boldsymbol{x}^{(k)\top} \boldsymbol{\Lambda}^2 \boldsymbol{x}^{(k)} \right)^2}{\left(\boldsymbol{x}^{(k)\top} \boldsymbol{\Lambda}^3 \boldsymbol{x}^{(k)} \right) \left(\boldsymbol{x}^{(k)\top} \boldsymbol{\Lambda} \boldsymbol{x}^{(k)} \right)} \right) f(\boldsymbol{x}^{(k)}) \\
&\leq \left(1 - \frac{4\lambda_1 \lambda_d}{(\lambda_1 + \lambda_d)^2} \right) f(\boldsymbol{x}^{(k)}) = \left(\frac{\lambda_1 - \lambda_d}{\lambda_1 + \lambda_d} \right)^2 f(\boldsymbol{x}^{(k)})
\end{aligned}$$

The last inequality uses Kantorovich Inequality

$$\frac{\left(\boldsymbol{y}^\top \boldsymbol{\Lambda}^2 \boldsymbol{y} \right)^2}{\left(\boldsymbol{y} \boldsymbol{\Lambda}^3 \boldsymbol{y} \right) \left(\boldsymbol{y}^\top \boldsymbol{\Lambda} \boldsymbol{y} \right)} \geq \frac{4\lambda_1 \lambda_d}{(\lambda_1 + \lambda_d)^2}.$$

11.7.4 Adaptive Step Size Momentum Method

For asymptotic analysis, we consider the region near the optimum, in which the objective function can be well-approximated by a quadratic. We know that fixing $\alpha^{(k)}$ to $\frac{2}{\lambda_1 + \lambda_d}$ yields

$$\left\| \boldsymbol{y}^{(k+1)} \right\| \leq \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \left\| \boldsymbol{y}^{(k)} \right\|.$$

On the other hand, choosing adaptive step size

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} \nabla f^\top \nabla^2 f \nabla f & -\Delta \boldsymbol{x}^\top \nabla^2 f \nabla f \\ -\Delta \boldsymbol{x}^\top \nabla^2 f \nabla f & \Delta \boldsymbol{x}^\top \nabla^2 f \Delta \boldsymbol{x} \end{bmatrix}^{-1} \begin{bmatrix} \nabla f^\top \nabla f \\ -\Delta \boldsymbol{x}^\top \nabla f \end{bmatrix}$$

minimizes the quadratic with respect to α, β. That means the resulting $\hat{\boldsymbol{y}}^{(k)}$ satisfies

$$\left\| \hat{\boldsymbol{y}}^{(k+1)} \right\| \leq \left\| \boldsymbol{y}^{(k+1)} \right\| \leq \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \left\| \boldsymbol{y}^{(k)} \right\|.$$

Hence, each iteration of adaptive schedule decreases the distance at least as much as each iteration of fixed step size scheme. The convergence rate therefore is upper-bounded by the one of fixed step size scheme inside the quadratic region, i.e., $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.

Chapter 12: Conclusion and Future Work

In summary, the contributions of this dissertation are as follows:

- A closed-form bound on the convergence of iterative methods via fixed point analysis is developed in Chapter 2. This enables the establishment of the convergence rate, region of convergence, and the number of required iterations to reach certain accuracy.

- A unified framework to study the local linear convergence of projected gradient descent for constrained least squares is proposed in Chapter 3. The proposed framework relies on three key steps: *(i)* the introduction of Lipschitz-continuous differentiability to provide tight error bounds on the linear approximation of the projection operator near the solution, *(ii)* the establishment of an asymptotically-linear recursion on the error iterations, and *(iii)* the derivation of the linear rate and the region of convergence of the error sequence.

- Utilizing this unified framework, the convergence rate analysis of projected gradient descent for minimizing a quadratic over a sphere is developed in Chapter 4 and then the result to the unit-modulus constrained least squares problem is generalized in Chapter 5. In each problem, acceleration techniques is further proposed to improve the performance of the algorithm in practice.

- Next, the convergence analysis of iterative hard thresholding for matrix completion is established. Thanks to the aforementioned unified framework, convergence analysis is shifted to the characterization of the rank-r projection operator. The perturbation expansion and error bound for this projection are developed in Chapter 6. Then, the exact rate of convergence and its asymptotic behavior in large-scale matrix completion setting are analyzed in Chapter 7.

- In Chapters 8 and 9, two acceleration techniques for IHT that can be used to exploit the asymptotic convergence results and obtain the optimal convergence in practice are demonstrated.

- Chapter 10 analyzed gradient descent for the factorization-based formulation of matrix completion. Under the view of fixed-point iterations, the first known closed-form expression of the convergence rate is established.

- Finally, a practical algorithm for selecting momentum step sizes in deconvolution applications is proposed. The proposed method recovers the optimal rate of convergence for the Heavy-Ball method while remaining the same computational complexity as traditional gradient schemes.

Below is the list of the publications during my PhD study.

Journals:

1. **Trung Vu**, Raviv Raich, and Xiao Fu, "On Local Linear Convergence of Gradient Projection for Unit-Modulus Least Squares", arXiv preprint arXiv:2206.10832, 2022. *Under review.*

2. **Trung Vu**, Evgenia Chunikhina, and Raviv Raich, "On Asymptotic Linear Convergence Rate of Iterative Hard Thresholding for Matrix Completion", arXiv preprint arXiv:2112.14733, 2022. *Under review.*

3. **Trung Vu** and Raviv Raich, "On Asymptotic Linear Convergence of Projected Gradient Descent for Constrained Least Squares", IEEE Transactions on Signal Processing, vol. 70, pp. 4061-4076, 2022.

4. **Trung Vu** and Raviv Raich, "A Closed-Form Bound on the Asymptotic Linear Convergence of Iterative Methods via Fixed Point Analysis", Optimization Letters, vol. 1, pp. 1-14, 2022.

5. **Trung Vu**, Evgenia Chunikhina, and Raviv Raich, "Perturbation Expansions and Error Bounds for the Truncated Singular Value Decomposition", Linear Algebra and Its Applications, vol. 627, pp. 94-139, 2021.

6. **Trung Vu**, Phung Lai, Raviv Raich, Anh Pham, Xiaoli Z. Fern and UK Arvind Rao, "A Novel Attribute-based Symmetric Multiple Instance Learning for Histopathological Image Analysis", IEEE Transactions on Medical Imaging, vol. 39, no. 10, pp. 3125-3136, 2020.

Conference papers:

1. **Trung Vu** and Raviv Raich, "Exact Linear Convergence Rate Analysis for Low-Rank Symmetric Matrix Completion via Gradient Descent", In Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 3240-3244. IEEE, 2021.

2. **Trung Vu**, Raviv Raich, and Xiao Fu, "On Convergence of Projected Gradient Descent for Minimizing a Large-scale Quadratic over the Unit Sphere", In Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6. IEEE, 2019. **Received Best Student Paper Award.**

3. **Trung Vu** and Raviv Raich, "Local Convergence of the Heavy Ball method in Iterative Hard Thresholding for Low-Rank Matrix Completion", In Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 3417-3421. IEEE, 2019.

4. **Trung Vu** and Raviv Raich, "Accelerating Iterative Hard Thresholding for Low-Rank Matrix Completion via Adaptive Restart", In Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 2917-2921. IEEE, 2019.

5. **Trung Vu**, Raviv Raich, "Adaptive Step Size Momentum Method For Deconvolution", In Proceedings of IEEE Statistical Signal Processing Workshop (SSP), pp. 438-442. IEEE, 2018.

The convergence analysis framework presented in this dissertation can be used as a general recipe to develop quick yet sharp local convergence results for iterative algorithms in various MLSP applications as well as to complement existing analyses of convergence. In the following, directions for future research in convergence rate analysis are highlighted.

12.1 Asymptotic Convergence Rate for Low-Rank Asymmetric Matrix Completion via Factorized-Based Gradient Descent

The asymmetric matrix completion problem is a more general version of symmetric matrix completion [215]. It can be formulated as

$$f(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{2} \left\| \mathcal{P}_{\Omega}(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M}) \right\|_F^2$$

The derivatives of $f(\boldsymbol{X})$ with respect to $\boldsymbol{X}$ and $\boldsymbol{Y}$ are

$$\begin{aligned} \nabla_{\boldsymbol{X}} f(\boldsymbol{X}, \boldsymbol{Y}) &= \mathcal{P}_{\Omega}(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M})\boldsymbol{X} \\ \text{and } \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}, \boldsymbol{Y}) &= \mathcal{P}_{\Omega}(\boldsymbol{X}\boldsymbol{Y}^{\top} - \boldsymbol{M})^{\top}\boldsymbol{Y}. \end{aligned}$$

The PGD update is given by

$$\begin{bmatrix} \boldsymbol{X}^{(k+1)} \\ \boldsymbol{Y}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{(k)} \\ \boldsymbol{Y}^{(k)} \end{bmatrix} - \eta \begin{bmatrix} \mathcal{P}_{\Omega}(\boldsymbol{X}^{(k)}(\boldsymbol{Y}^{(k)})^{\top} - \boldsymbol{M})\boldsymbol{X}^{(k)} \\ \mathcal{P}_{\Omega}(\boldsymbol{X}^{(k)}(\boldsymbol{Y}^{(k)})^{\top} - \boldsymbol{M})^{\top}\boldsymbol{Y}^{(k)} \end{bmatrix}.$$

For the non-symmetric case, instead of considering a direct recursion on the error matrix $\boldsymbol{E}^k = \boldsymbol{X}^k(\boldsymbol{X}^k)^{\top} - \boldsymbol{M}$, we need to construct the recursive equations on two separate error matrices, i.e., $[\boldsymbol{\Delta}X^{k+1}, \boldsymbol{\Delta}Y^{k+1}] = g([\boldsymbol{\Delta}X^k, \boldsymbol{\Delta}Y^k])$. The next challenge is to find a first-order expansion of the function g that can approximate it well in the neighborhood of the solution where the error matrices are small. Finally, one would determine the asymptotic linear convergence rate by analyzing

of the behavior of the linear operator under structural constraints of the problem (as we have seen in the symmetric case).

12.2 Minimum-Norm Adversarial Attacks using Gradient Projections with Spherical Constraints

Given an image $\boldsymbol{a} \in \mathbb{R}^n$ that belongs to class c. The problem of finding an adversarial instance for $\boldsymbol{a}$ can be formulated as

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \|\boldsymbol{x} - \boldsymbol{a}\| \qquad \text{subject to } C(\boldsymbol{x}) = t, \tag{12.1}$$

where $C(\boldsymbol{x})$ is the output label assigned by the classifier for image $\boldsymbol{x}$ and $t \neq c$ is the target class. Since the constraint $C(\boldsymbol{x}) = t$ is often highly non-linear (e.g., neural networks), we consider an objective function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\begin{cases} f(\boldsymbol{x}) > 0 & \text{if } C(\boldsymbol{x}) \neq t, \\ f(\boldsymbol{x}) \leq 0 & \text{if } C(\boldsymbol{x}) = t. \end{cases}$$

Thus, problem (12.1) can be reformulated as [38]

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \|\boldsymbol{x} - \boldsymbol{a}\| \qquad \text{subject to } f(\boldsymbol{x}) \leq 0. \tag{12.2}$$

The above problem can be solved via iterative expanding radius approach as follows. Starting from a sufficiently small radius $R = R_0$, the algorithm works in a

way that it alternates between (i) minimizing the objective function on the sphere with center $\boldsymbol{a}$ and radius R and (ii) increasing the radius by an appropriate amount. The iterative process stops when we find $\boldsymbol{x}$ such that $f(\boldsymbol{x}) < 0$. Thus, our optimization problem is essentially reduced to solving a sequence of optimization problems:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \qquad \text{subject to } \|\boldsymbol{x}-\boldsymbol{a}\| = R_t, \tag{12.3}$$

where R_t is the radius at the t-th iteration. In order to solve the subproblem (12.3), we use the so-called projected gradient descent method (PGD) with fixed step size, where the projection onto the sphere with center $\boldsymbol{a}$ and radius R is defined as

$$\mathcal{P}_{\boldsymbol{a},R}(\boldsymbol{x}) = \begin{cases} \boldsymbol{a} + R\frac{\boldsymbol{x}-\boldsymbol{a}}{\|\boldsymbol{x}-\boldsymbol{a}\|} & \text{if } \boldsymbol{x} \neq \boldsymbol{a}, \\ R\boldsymbol{e}_1 & \text{if } \boldsymbol{x} = \boldsymbol{a}. \end{cases} \tag{12.4}$$

One advantage of this approach is that it allows to trace the trajectory of optimal solutions as the radius increase. By keeping the iterates close to the optimal trajectory, one may hope that the algorithm will be able to find a good solution at the end of the path. Given the analysis we performed with spherical constrained quadratic minimization problems [218], it would be interesting to propose such efficient algorithm with theoretical guarantees. Since f is highly nonlinear and potentially non-smooth, there are several challenges with the analysis here including the guarantee on the convergence to stationary points and the strategy to expand the radius so that the iterates remains close to the optimal trajectory.

12.3 Other Long-Term Research Directions

Below are some topics that are of interest in this dissertation but require more time and effort to work on. They can be classified as long-term goals.

1. **Derivative of the projection onto manifolds using shape operators**: The gradient of the projection operator can be further computed using recent results in differential geometry on the shape operator. It would be interesting to see how it can be applied to our expression of the asymptotic rate.

2. **Matrix Completion in different asymptotic regimes**: In most of recent works on the global convergence of algorithms for low-rank matrix completion, the setting of interest is the asymptotic case when the rank is constant and the number of observations grows almost linear in the dimension of the matrix. It would be interesting to make connections between our local convergence result and the existing global convergence result, e.g., the selection of the step size, the region of convergence, the rate of convergence, how random matrix theory can be applied in such regime.

3. **Convergence of PGD with backtracking and/or exact line search**: It has been well-known that PGD with exact line search obtains the optimal rate of convergence. Interestingly, this coincides with the optimal fixed step size in our aforementioned analysis. It is interesting to study why this happens and how our analysis for the fixed step size scheme can extended to the schemes with adaptive step size.

Bibliography

[1] Mohammed M Abo-Zahhad, Aziza I Hussein, Abdelfatah M Mohamed, et al. Compression of ECG signal based on compressive sensing and the extraction of significant features. *International Journal of Communications, Network and System Sciences*, 8(05):97, 2015.

[2] Milton Abramowitz and Irene A Stegun. Handbook of mathematical functions with formulas, graphs, and mathematical tables. *NBS Applied Mathematics Series*, 55, 1964.

[3] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[4] P-A Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

[5] Zeyuan Allen-Zhu and Yuanzhi Li. LazySVD: Even faster SVD decomposition yet without agonizing pain. *Proceedings of Advances in Neural Information Processing Systems*, 29:974–982, 2016.

[6] Luigi Ambrosio and Carlo Mantegazza. Curvature and distance function from a manifold. *Journal of Geometric Analysis*, 8(5):723–748, 1998.

[7] Isaac Amidror. Scattered data interpolation methods for electronic imaging systems: A survey. *Journal of Electronic Imaging*, 11(2):157–176, 2002.

[8] Harry C Andrews and Boby Ray Hunt. *Digital image restoration*. Prentice-Hall, 1977.

[9] Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

[10] Mark R Banham and Aggelos K Katsaggelos. Digital image restoration. *IEEE Signal Processing Magazine*, 14(2):24–41, 1997.

[11] Buyurmaiz Baykal. Blind channel estimation via combining autocorrelation and blind phase estimation. *IEEE Transactions on Circuits and Systems—Part I: Regular Papers*, 51(6):1125–1131, 2004.

[12] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[13] Amir Beck and Yakov Vaisbourd. Globally solving the trust region subproblem using simple first-order methods. *SIAM Journal on Optimization*, 28(3):1951–1967, 2018.

[14] Richard Bellman. *Stability theory of differential equations*. McGraw-Hill, NY, New York, 1953.

[15] Richard Bellman. *Stability theory of differential equations*. McGraw-Hill, NY, New York, 1953.

[16] Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[17] Dimitri P Bertsekas and Eli M Gafni. Projection methods for variational inequalities with application to the traffic assignment problem. In *Nondifferential and variational techniques in optimization*, pages 139–159. Springer, 1982.

[18] Jeffrey D Blanchard, Jared Tanner, and Ke Wei. CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion. *Information and Inference: A Journal of the IMA*, 4(4):289–327, 2015.

[19] William E Blass and George W Halsey. *Deconvolution of absorption spectra*. Academic Press, 1981.

[20] Thomas Blumensath and Mike E Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008.

[21] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[22] Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1):197–200, 1992.

[23] Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.

[24] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[25] Luitzen Egbertus Jan Brouwer. Über abbildung von mannigfaltigkeiten. *Mathematische Annalen*, 71(1):97–115, 1911.

[26] Samuel Burer and Renato DC Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[27] 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.

[28] Jian-Feng Cai, Tianming Wang, and Ke Wei. Spectral compressed sensing via projected gradient descent. *SIAM Journal on Optimization*, 28(3):2625–2653, 2018.

[29] T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.

[30] Emmanuel Candes and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969, 2007.

[31] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

[32] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[33] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.

[34] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[35] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[36] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[37] Mireille Capitaine and Muriel Casalis. Asymptotic freeness by generalized moments for gaussian and wishart matrices. Application to beta random matrices. *Indiana University Mathematics Journal*, pages 397–431, 2004.

[38] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proceedings of IEEE Symposium on Security and Privacy*, pages 39–57. IEEE, 2017.

[39] Rakesh Chalasani, Jose C Principe, and Naveen Ramakrishnan. A fast proximal method for convolutional sparse coding. In *Proceedings of International Joint Conference on Neural Networks*, pages 1–5. IEEE, 2013.

[40] Caihua Chen, Bingsheng He, and Xiaoming Yuan. Matrix completion via an alternating direction method. *IMA Journal of Numerical Analysis*, 32(1):227–245, 2012.

[41] Yangkang Chen, Jiang Yuan, Shaohuan Zu, Shan Qu, and Shuwei Gan. Seismic imaging of simultaneous-source data using constrained least-squares reverse time migration. *Journal of Applied Geophysics*, 114:32–35, 2015.

[42] Yudong Chen and Yuejie Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, 2018.

[43] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[44] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[45] Alexander L Chistov and D Yu Grigor'Ev. Complexity of quantifier elimination in the theory of algebraically closed fields. In *Proceedings of International Symposium on Mathematical Foundations of Computer Science*, pages 17–31. Springer, 1984.

[46] Evgenia Chunikhina, Raviv Raich, and Thinh Nguyen. Performance analysis for matrix completion via iterative hard-thresholded SVD. In *Proceedings of IEEE Statistical Signal Processing Workshop*, pages 392–395. IEEE, 2014.

[47] Benoît Collins. Product of random projections, Jacobi ensembles and universality problems arising from free probability. *Probability Theory and Related Fields*, 133(3):315–344, 2005.

[48] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[49] Andrew R Conn, Nicholas IM Gould, and Ph L Toint. *Trust region methods*, volume 1. SIAM, 2000.

[50] John P Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16(1):2859–2900, 2015.

[51] James W Daniel. The conjugate gradient method for linear and nonlinear operator equations. *SIAM Journal on Numerical Analysis*, 4(1):10–26, 1967.

[52] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[53] Marcello L.R. de Campos, Stefan Werner, and José A Apolinário. Constrained adaptive filters. In *Adaptive Antenna Arrays: Trends and Applications*, pages 46–64. Springer Berlin Heidelberg, 2004.

[54] Etienne De Klerk, François Glineur, and Adrien B Taylor. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11(7):1185–1199, 2017.

[55] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 66(11):7274–7301, 2020.

[56] Aleksandar Dogandžić, Renliang Gu, and Kun Qiu. Mask iterative hard thresholding algorithms for sparse image reconstruction of objects with known contour. In *Conference Record of the Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 2111–2116. IEEE, 2011.

[57] Zhishan Dong, Tiefeng Jiang, and Danning Li. Circular law and arc law for truncation of random unitary matrix. *Journal of Mathematical Physics*, 53(1):013301, 2012.

[58] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[59] Ewa Dudek and Konstanty Holly. Nonlinear orthogonal projection. *Annales Polonici Mathematici*, 59(1):1–31, 1994.

[60] Joseph C Dunn. Global and asymptotic convergence rate estimates for a class of projected gradient processes. *SIAM Journal on Control and Optimization*, 19(3):368–400, 1981.

[61] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[62] Alan Edelman and N Raj Rao. Random matrix theory. *Acta Numerica*, 14:233, 2005.

[63] Brendan Farrell and Raj Rao Nadakuditi. Local spectrum of truncations of Kronecker products of Haar distributed unitary matrices. *Random Matrices: Theory and Applications*, 4(1), 2013.

[64] Florian Feppon and Pierre FJ Lermusiaux. A geometric approach to dynamical model order reduction. *SIAM Journal on Matrix Analysis and Applications*, 39(1):510–538, 2018.

[65] Mário AT Figueiredo and Robert D Nowak. A bound optimization approach to wavelet-based image deconvolution. In *Proceedings of IEEE International Conference on Image Processing*, volume 2, pages II–782. IEEE, 2005.

[66] Mário AT Figueiredo, Robert D Nowak, and Stephen J Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007.

[67] Robert L Foote. Regularity of the distance function. *Proceedings of American Mathematical Society*, 92(1):153–155, 1984.

[68] Peter J Forrester. Quantum conductance problems and the Jacobi ensemble. *Journal of Physics A: Mathematical and General*, 39(22):6861, 2006.

[69] Otis Lamont Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of IEEE*, 60(8):926–935, 1972.

[70] Xiao Fu, Frankie KW Chan, Wing-Kin Ma, and HC So. A complex-valued semidefinite relaxation approach for two-dimensional source localization using distance measurements and imperfect receiver positions. In *IEEE International Conference on Signal Processing*, volume 2, pages 1491–1494. IEEE, 2012.

[71] Daniel Gabay. Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, 37(2):177–219, 1982.

[72] Nikolas P Galatsanos, Aggelos K Katsaggelos, Roland T Chin, and Allen D Hillery. Least squares restoration of multichannel images. *IEEE Transactions on Signal Processing*, 39(10):2222–2236, 1991.

[73] Jean Gallier. Quadratic optimization and contour grouping. In *Geometric Methods and Applications*, pages 439–457. Springer, 2011.

[74] David Yang Gao. *Duality principles in nonconvex systems: theory, methods and applications*, volume 39. Springer Science & Business Media, 2013.

[75] Zhen Gao, Chao Zhang, Zhaocheng Wang, and Sheng Chen. Priori-information aided iterative hard threshold: A low-complexity high-accuracy compressive sensing based channel estimation for TDS-OFDM. *IEEE Transactions on Wireless Communications*, 14(1):242–251, 2014.

[76] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[77] Izrail Gelfand. Normierte ringe. *Matematicheskii Sbornik*, 9(1):3–24, 1941.

[78] Ralph W Gerchberg. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[79] Donald Goldfarb and Shiqian Ma. Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics*, 11(2):183–210, 2011.

[80] Gene H Golub and Urs Von Matt. Quadratically constrained least squares and quadratic problems. *Numerische Mathematik*, 59(1):561–580, 1991.

[81] Alexander Graham. *Kronecker products and matrix calculus with applications*. Courier Dover Publications, 2018.

[82] Serge Gratton and Jean Tshimanga. On a second-order expansion of the truncated singular subspace decomposition. *Numerical Linear Algebra with Applications*, 23(3):519–534, 2016.

[83] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan. NeNMF: An optimal gradient method for nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 60(6):2882–2898, 2012.

[84] Bahadir K Gunturk, Yucel Altunbasak, and Russell M Mersereau. Color plane interpolation using alternating projections. *IEEE Transactions on Image Processing*, 11(9):997–1013, 2002.

[85] Qiang Guo, Caiming Zhang, Yunfeng Zhang, and Hui Liu. An efficient SVD-based method for image denoising. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(5):868–880, 2015.

[86] William Hager and Soonchul Park. Global convergence of SSM for minimizing a quadratic over a sphere. *Mathematics of Computation*, 74(251):1413–1423, 2005.

[87] William W Hager. Minimizing a quadratic over a sphere. *SIAM Journal on Optimization*, 12(1):188–208, 2001.

[88] William W Hager and Yaroslav Krylyuk. Graph partitioning and continuous quadratic programming. *SIAM Journal on Discrete Mathematics*, 12(4):500–523, 1999.

[89] Per Christian Hansen. The truncated SVD as a method for regularization. *BIT Numerical Mathematics*, 27(4):534–553, 1987.

[90] Per Christian Hansen. Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank. *SIAM Journal on Scientific and Statistical Computing*, 11(3):503–518, 1990.

[91] Moritz Hardt. Understanding alternating minimization for matrix completion. In *Proceedings of Annual IEEE Symposium on Foundations of Computer Science*, pages 651–660, 2014.

[92] Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. In *Proceedings of Conference on Learning Theory*, pages 638–678, 2014.

[93] Godfrey Harold Hardy, John Edensor Littlewood, George Pólya, and György Pólya. *Inequalities*. Cambridge Mathematical Library. Cambridge University Press, 1952.

[94] Adrian Hauswirth, Saverio Bolognani, Gabriela Hug, and Florian Dörfler. Projected gradient descent on Riemannian manifolds with applications to online power system optimization. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 225–232. IEEE, 2016.

[95] A Hedayat, Walter Dennis Wallis, et al. Hadamard matrices and their applications. *The Annals of Statistics*, 6(6):1184–1238, 1978.

[96] Xiaoying Hong, Xiaoping Lai, and Ruijie Zhao. Matrix-based algorithms for constrained least-squares and minimax designs of 2-d linear-phase FIR filters. *IEEE Transactions on Signal Processing*, 61(14):3620–3631, 2013.

[97] Erik Hons. *Constrained quadratic optimization: Theory and application for wireless communication systems*. PhD thesis, University of Waterloo, 2001.

[98] Wei Huang, Liang Xiao, Hongyi Liu, and Zhihui Wei. Hyperspectral imagery super-resolution by compressive sensing inspired dictionary learning and spatial-spectral regularization. *Sensors*, 15(1):2041–2058, 2015.

[99] B Hunt. A matrix theory proof of the discrete convolution theorem. *IEEE Transactions on Audio and Electroacoustics*, 19(4):285–288, 1971.

[100] Bobby R Hunt. The application of constrained least squares estimation to image restoration by digital computer. *IEEE Transactions on Computers*, 100(9):805–812, 1973.

[101] Suk-Geun Hwang. Cauchy's interlace theorem for eigenvalues of Hermitian matrices. *The American Mathematical Monthly*, 111(2):157–159, 2004.

[102] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

[103] Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3-4):142–336, 2017.

[104] Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Proceedings of Advances in Neural Information Processing Systems*, pages 937–945, 2010.

[105] Prateek Jain and Praneeth Netrapalli. Fast exact matrix completion with finite samples. In *Proceedings of Conference on Learning Theory*, pages 1007–1034, 2015.

[106] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of Annual ACM Symposium on Theory of Computing*, pages 665–674, 2013.

[107] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proceedings International Conference on Machine Learning*, pages 457–464, 2009.

[108] Tiefeng Jiang. Approximation of haar distributed matrices and limiting distributions of eigenvalues of Jacobi ensembles. *Probability Theory and Related Fields*, 144(1-2):221–246, 2009.

[109] Yindi Jing and Xinwei Yu. ML-based channel estimations for non-regenerative relay networks with multiple transmit and receive antennas. *IEEE Journal on Selected Areas in Communications*, 30(8):1428–1439, 2012.

[110] Björn Johansson, Tommy Elfving, Vladimir Kozlov, Yair Censor, Per-Erik Forssén, and Gösta Granlund. The application of an oblique-projected

Landweber method to a model of supervised learning. *Mathematical and Computer Modelling*, 43(7-8):892–909, 2006.

[111] Iain M Johnstone. Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *The Annals of Statistics*, 36(6):2638, 2008.

[112] Nocedal Jorge and J Wright Stephen. *Numerical optimization*. Spinger, 2006.

[113] Alexander Jung. A fixed-point of view on gradient methods for big data. *Frontiers in Applied Mathematics and Statistics*, 3:18, 2017.

[114] Leonid Vitalévich Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6):89–185, 1948.

[115] Raghunandan Hulikal Keshavan. *Efficient algorithms for collaborative filtering*. PhD thesis, Stanford University, 2012.

[116] Rajiv Khanna and Anastasios Kyrillidis. IHT dies hard: Provable accelerated iterative hard thresholding. In *Proceedings of International Conference on Artificial Intelligence Statistics*, pages 188–198, 2018.

[117] Yehuda Koren. The BellKor solution to the Netflix grand prize. *Netflix prize documentation*, 81(2009):1–10, 2009.

[118] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[119] Anastasios Kyrillidis and Volkan Cevher. Matrix recipes for hard thresholding methods. *Journal of Mathematical Imaging and Vision*, 48(2):235–265, 2014.

[120] Ming Jun Lai and Abraham Varghese. On convergence of the alternating projection method for matrix completion and sparse recovery problems. *arXiv preprint arXiv:1711.02151*, 2017.

[121] James V Lambers, Amber Sumner Mooney, and Vivian A Montiforte. *Explorations in numerical analysis*. World Scientific, 2019.

[122] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. *Proceedings of Advances in Neural Information Processing Systems*, 19, 2006.

[123] John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer, 2013.

[124] John M Lee. *Introduction to Riemannian manifolds*. Springer, 2018.

[125] Kiryung Lee and Yoram Bresler. ADMiRA: Atomic decomposition for minimum rank approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416, 2010.

[126] M Lee John. Introduction to smooth manifolds. *Graduate Texts in Mathematics*, 218, 2003.

[127] Gunther Leobacher and Alexander Steinicke. Existence, uniqueness and regularity of the projection onto differentiable manifolds. *Annals of Global Analysis and Geometry*, pages 1–29, 2021.

[128] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[129] Adrian S Lewis and Jérôme Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234, 2008.

[130] Fu Li and Richard J Vaccaro. Unified analysis for DOA estimation algorithms in array signal processing. *Signal Processing*, 25(2):147–169, 1991.

[131] Zi-Cai Li, Hung-Tsai Huang, and Yimin Wei. Ill-conditioning of the truncated singular value decomposition, Tikhonov regularization and their applications to numerical partial differential equations. *Numerical Linear Algebra with Applications*, 18(2):205–221, 2011.

[132] A Lichnewsky. *Minimisation des Fonctionnelles Définies sur une Variété par la Methode du Gradiënt Conjugué*. PhD thesis, These de Doctorat d'Etat. Paris: Université de Paris-Sud, 1979.

[133] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[134] Erik Lindskog and Claes Tidestav. Reduced rank channel estimation. In *IEEE Vehicular Technology Conference*, volume 2, pages 1126–1130, 1999.

[135] Jun Liu, Xiangqian Liu, and Xiaoli Ma. First-order perturbation analysis of singular vectors in singular value decomposition. *IEEE Transactions on Signal Processing*, 56(7):3044–3049, 2008.

[136] Zhang Liu, Anders Hansson, and Lieven Vandenberghe. Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters*, 62(8):605–612, 2013.

[137] 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, 2010.

[138] Cheng-Jun Lu, Wei-Xing Sheng, Yu-Bing Han, and Xiao-Feng Ma. A novel adaptive phase-only beamforming algorithm based on semidefinite relaxation. In *IEEE International Symposium on Phased Array Systems and Technology*, pages 617–621. IEEE, 2013.

[139] David G Luenberger. The gradient projection method along geodesics. *Management Science*, 18(11):620–631, 1972.

[140] David G Luenberger, Yinyu Ye, et al. *Linear and nonlinear programming*, volume 2. Springer, 1984.

[141] Zhi-Quan Luo, Wing-Kin Ma, Anthony Man-Cho So, Yinyu Ye, and Shuzhong Zhang. Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Processing Magazine*, 27(3):20–34, 2010.

[142] Zhi-Quan Luo, Nicholas D Sidiropoulos, Paul Tseng, and Shuzhong Zhang. Approximation bounds for quadratic optimization with homogeneous quadratic constraints. *SIAM Journal on Optimization*, 18(1):1–28, 2007.

[143] Zhi-Quan Luo and Paul Tseng. Error bound and reduced-gradient projection algorithms for convex minimization over a polyhedral set. *SIAM Journal on Optimization*, 3(1):43–59, 1993.

[144] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. In *Foundations of Computational Mathematics*, pages 451—-632, 2020.

[145] Shiqian Ma, Donald Goldfarb, and Lifeng Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353, 2011.

[146] Ivan Markovsky. Structured low-rank approximation and its applications. *Automatica*, 44(4):891–909, 2008.

[147] Ivan Markovsky, Jan C Willems, Sabine Van Huffel, Bart De Moor, and Rik Pintelon. Application of structured total least squares for system identification and model reduction. *IEEE Transactions on Automatic Control*, 50(10):1490–1500, 2005.

[148] Elaine Crespo Marques, Nilson Maciel, Lirida Naviner, Hao Cai, and Jun Yang. A review of sparse recovery algorithms. *IEEE Access*, 7:1300–1322, 2018.

[149] William Menke. *Geophysical data analysis: Discrete inverse theory*. Academic press, 2018.

[150] Vladimir Z Mesarovic, Nikolas P Galatsanos, and Aggelos K Katsaggelos. Regularized constrained total least squares image restoration. *IEEE Transactions on Image Processing*, 4(8):1096–1108, 1995.

[151] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. SIAM, 2000.

[152] Rick P Millane. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.

[153] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

[154] Nasser Mohammadiha and Arne Leijon. Nonnegative matrix factorization using projected gradient algorithms with sparseness constraints. In *IEEE Symposium on Signal Processing and Information Technology*, pages 418–423. IEEE, 2009.

[155] Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2140–2151, 2013.

[156] Karthik Mohan and Maryam Fazel. Reweighted nuclear norm minimization with application to system identification. In *Proceedings of American Control Conference*, pages 2953–2959. IEEE, 2010.

[157] M Nashed. Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory. *IEEE Transactions on Antennas and Propagation*, 29(2):220–231, 1981.

[158] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

[159] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

[160] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[161] Heinz Neudecker and Tom Wansbeek. Fourth-order properties of normally distributed random matrices. *Linear Algebra and its Applications*, 97:13–21, 1987.

[162] Monica Nicoli and Umberto Spagnolini. Reduced-rank channel estimation for time-slotted mobile communication systems. *IEEE Transactions on Signal Processing*, 53(3):926–944, 2005.

[163] Ricardo Otazo, Emmanuel Candes, and Daniel K Sodickson. Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components. *Magnetic Resonance in Medicine*, 73(3):1125–1136, 2015.

[164] Brendan O'donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

[165] Jaehyun Park and Stephen Boyd. General heuristics for nonconvex quadratically constrained quadratic programming. *arXiv preprint arXiv:1703.07870*, 2017.

[166] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[167] Boris T Polyak. *Introduction to optimization. Optimization software*, volume 1. Inc., Publications Division, New York, 1987.

[168] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963.

[169] Charles I Puryear, Oleg N Portniaguine, Carlos M Cobos, and John P Castagna. Constrained least-squares spectral analysis: Application to seismic data. *Geophysics*, 77(5):V143–V167, 2012.

[170] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice Hall, 1993.

[171] Raviv Raich and Jinsub Kim. On the eigenvalue distribution of column subsampled semi-unitary matrices. In *Proceedings of IEEE Statistical Signal Processing Workshop*, pages 1–5. IEEE, 2016.

[172] 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.

[173] Jan Rataj and Martina Zähle. *Curvature measures of singular sets*. Springer, 2019.

[174] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

[175] Franz Rendl and Henry Wolkowicz. A semidefinite framework for trust region subproblems with applications to large scale minimization. *Mathematical Programming*, 77(1):273–299, 1997.

[176] Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings International Conference on Machine Learning*, pages 713–719. ACM, 2005.

[177] Leonardo S Resende, Joao Marcos T Romano, and Maurice G Bellanger. A fast least-squares algorithm for linearly constrained adaptive filtering. *IEEE Transactions on Signal Processing*, 44(5):1168–1174, 1996.

[178] AW Roberts. The derivative as a linear transformation. *The American Mathematical Monthly*, 76(6):632–638, 1969.

[179] Marielba Rojas. *A large-scale trust-region approach to the regularization of discrete ill-posed problems*. PhD thesis, Rice University, 1999.

[180] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: Extreme singular values. In *Proceedings of the International Congress of Mathematicians*, volume 1, pages 1576–1602. World Scientific, 2010.

[181] R Saigal and Michael J Todd. Efficient acceleration techniques for fixed point algorithms. *SIAM Journal on Numerical Analysis*, 15(5):997–1007, 1978.

[182] Andrey A Shabalin and Andrew B Nobel. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, 2013.

[183] Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015.

[184] Mojtaba Soltanalian and Petre Stoica. Designing unimodular codes via quadratic optimization. *IEEE Transactions on Signal Processing*, 62(5):1221–1234, 2014.

[185] Michal Šorel and Filip Šroubek. Fast convolutional sparse coding using matrix inversion lemma. *Digital Signal Processing*, 55:44–51, 2016.

[186] Danny C Sorensen. Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19(2):409–426, 1982.

[187] Danny C Sorensen. Minimization of a large-scale quadratic function subject to a spherical constraint. *SIAM Journal on Optimization*, 7(1):141–161, 1997.

[188] Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *Proceedings International Conference on Machine Learning*, pages 720–727, 2003.

[189] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1329–1336, 2005.

[190] Gilbert W Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Review*, 15(4):727–764, 1973.

[191] Gilbert W Stewart. A second order perturbation expansion for small singular values. *Linear Algebra and its Applications*, 56:231–235, 1984.

[192] Gilbert W Stewart. Perturbation theory for the singular value decomposition. In *SVD and Signal Processing, II: Algorithms, Analysis and Applications*, pages 99–109. Elsevier, 1990.

[193] Gilbert W Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[194] Michael Stewart. Perturbation of the SVD in the presence of small singular values. *Linear Algebra and its Applications*, 419(1):53–77, 2006.

[195] Christoph Stöckle, Jawad Munir, Amine Mezghani, and Josef A Nossek. Channel estimation in massive MIMO systems using 1-bit quantization. In *IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–6. IEEE, 2016.

[196] Christoph Stoeckle, Jawad Munir, Amine Mezghani, and Josef A Nossek. DOA estimation performance and computational complexity of subspace- and compressed sensing-based methods. In *International ITG Workshop on Smart Antennas*, pages 1–6. VDE, 2015.

[197] W Gilbert Strang. On the Kantorovich inequality. *Proceedings of the American Mathematical Society*, 11(468):0095–09601, 1960.

[198] Ji-Guang Sun. A note on simple non-zero singular values. *Journal of Computational Mathematics*, pages 258–266, 1988.

[199] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[200] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Investigation of various matrix factorization methods for large recommender systems. In *Proceedings of IEEE International Conference on Data Mining Workshop*, pages 553–562. IEEE, 2008.

[201] Jared Tanner and Ke Wei. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

[202] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

[203] P Thompson. Adaptation by direct phase-shift adjustment in narrow-band adaptive antenna systems. *IEEE Transactions on Antennas and Propagation*, 24(5):756–760, 1976.

[204] Andreas M Tillmann and Marc E Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2013.

[205] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.

[206] John Tranter, Nicholas D Sidiropoulos, Xiao Fu, and Ananthram Swami. Fast unit-modulus least squares with applications in beamforming. *IEEE Transactions on Signal Processing*, 65(11):2875–2887, 2017.

[207] André Uschmajew and Bart Vandereycken. Geometric methods on low-rank matrix and tensor manifolds. In *Variational methods for nonlinear geometric data and applications*, pages 261–313. Springer Nature, 2020.

[208] Richard J Vaccaro. A second-order perturbation expansion for the SVD. *SIAM Journal on Matrix Analysis and Applications*, 15(2):661–671, 1994.

[209] Charles Van Loan. On the method of weighting for equality-constrained least-squares problems. *SIAM Journal on Numerical Analysis*, 22(5):851–864, 1985.

[210] FP Vasilyev and A Yu Ivanitskiy. *In-depth analysis of linear programming*. Springer Science & Business Media, 2013.

[211] Trung Vu, Evgenia Chunikhina, and Raviv Raich. Perturbation expansions and error bounds for the truncated singular value decomposition. *Linear Algebra and its Applications*, 627:94–139, 2021.

[212] Trung Vu and Raviv Raich. Adaptive step size momentum method for deconvolution. In *Proceedings of IEEE Statistical Signal Processing Workshop*, pages 438–442, 2018.

[213] Trung Vu and Raviv Raich. Accelerating iterative hard thresholding for low-rank matrix completion via adaptive restart. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2917–2921. IEEE, 2019.

[214] Trung Vu and Raviv Raich. Local convergence of the Heavy Ball method in iterative hard thresholding for low-rank matrix completion. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3417–3421. IEEE, 2019.

[215] Trung Vu and Raviv Raich. Exact linear convergence rate analysis for low-rank symmetric matrix completion via gradient descent. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3240–3244. IEEE, 2021.

[216] Trung Vu and Raviv Raich. A closed-form bound on the asymptotic linear convergence of iterative methods via fixed point analysis. *Optim. Lett.*, 1:1–14, 2022.

[217] Trung Vu and Raviv Raich. On asymptotic linear convergence of projected gradient descent for constrained least squares. *IEEE Transactions on Signal Processing*, 4:4061–4076, 2022.

[218] Trung Vu, Raviv Raich, and Xiao Fu. On convergence of projected gradient descent for minimizing a large-scale quadratic over the unit sphere. In *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2019.

[219] Kenneth W Wachter. The limiting empirical measure of multiple discriminant ratios. *The Annals of Statistics*, 8:937–957, 1980.

[220] Irene Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, MaxCut and complex semidefinite programming. *Mathematical Programming*, 149(1):47–81, 2015.

[221] Homer F Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

[222] Adriaan Walther. The question of phase retrieval in optics. *Optica Acta: International Journal of Optics*, 10(1):41–49, 1963.

[223] Boying Wang and Fuzhen Zhang. Some inequalities for the eigenvalues of the product of positive semidefinite hermitian matrices. *Linear Algebra and its Applications*, 160:113–118, 1992.

[224] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[225] Jyh-Jong Wei, Chuang-Jan Chang, Nai-Kuan Chou, and Gwo-Jen Jan. ECG data compression using truncated singular value decomposition. *IEEE Transactions on Information Technology in Biomedicine*, 5(4):290–299, 2001.

[226] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

[227] Norbert Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964.

[228] John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2080–2088, 2009.

[229] Li Xu, Jimmy S Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. *Proceedings of Advances in Neural Information Processing Systems*, 27, 2014.

[230] Zhengyuan Xu. Perturbation analysis for subspace decomposition with applications in subspace-based algorithms. *IEEE Transactions on Signal Processing*, 50(11):2820–2830, 2002.

[231] Dan Yang, Zongming Ma, and Andreas Buja. Rate optimal denoising of simultaneously sparse and low rank matrices. *Journal of Machine Learning Research*, 17(1):3163–3189, 2016.

[232] Pavel Yaskov. The universality principle for spectral distributions of sample covariance matrices. *arXiv preprint arXiv:1410.5190*, 2014.

[233] Zeyu You, Raviv Raich, Xiaoli Z Fern, and Jinsub Kim. Discriminative recurring signal detection and localization. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2377–2381. IEEE, 2017.

[234] Yuanlong Yu, Zhenzhen Sun, Wenxing Zhu, and Jason Gu. A homotopy iterative hard thresholding algorithm with extreme learning machine for scene recognition. *IEEE Access*, 6:30424–30436, 2018.

[235] Lei Zhang, Mengdao Xing, Cheng-Wei Qiu, Jun Li, Jialian Sheng, Yachao Li, and Zheng Bao. Resolution enhancement for inversed synthetic aperture radar imaging under low SNR via improved compressive sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 48(10):3824–3838, 2010.

[236] Ming Zhang, Jianxing Li, Shitao Zhu, and Xiaoming Chen. Fast and simple gradient projection algorithms for phase-only beamforming. *IEEE Transactions on Vehicular Technology*, 2021.

[237] Shuzhong Zhang and Yongwei Huang. Complex quadratic optimization and semidefinite programming. *SIAM Journal on Optimization*, 16(3):871–890, 2006.

[238] Ilan Ziskind and Mati Wax. Maximum likelihood localization of multiple sources by alternating projection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(10):1553–1560, 1988.

[239] Karol Zyczkowski and Hans-Jürgen Sommers. Truncations of random unitary matrices. *Journal of Physics A: Mathematical and General*, 33(10):2045, 2000.