Low-Rank Matrix Recovery:
Manifold Geometry and Global Convergence

Thesis by
Ziyun Zhang

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended May 8, 2023

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Thomas Yizhao Hou, for his unyielding support throughout my years at Caltech. With his patience and encouragement, he turned my treacherous PhD journey into a truly rewarding and enjoyable experience. He gave me much freedom to explore my own interests, while insisting on the importance of the big picture. It is with his help that I grew into an independent researcher. His views toward career and life have a profound impact on me.

I am grateful to my thesis committee chair, Prof. Andrew Stuart, and my thesis committee members, Prof. Houman Owhadi and Prof. Joel Tropp. Their constructive criticism was very helpful to the improvement of this thesis. I have taken classes with each of them, where I learned both the immense beauty and the ultimate rigor of mathematics. I am also grateful to Prof. Steven Low for serving on my candidacy committee and helping me shape the early directions of this thesis.

I am deeply indebted to Dr. Zhenzhen Li, with whom I collaborated extensively on my thesis research. It is Zhenzhen who first introduced me to the low-rank recovery problems. Her vivid ideas and deep passion for research have always been an inspiration to me. I am also grateful for the discussion with many people in the Caltech CMS department, including but not limited to Dr. De Huang, Dr. Ka Chun Lam, Dr. Jiajie Chen, Dr. Shumao Zhang, Dr. Yifan Chen, Yixuan Wang, Changhe Yang, Dr. Franca Hoffmann, Eitan Levin, and many others.

I would like to express my appreciation for all the administrative staff in the CMS department, especially Diana Bohler, who made my PhD life much smoother by taking care of all administrative matters with great detail. I would like to thank NSF, Kaplun fellowship, and the Choi family gift fund for funding my research.

My PhD life is made memorable by many dear friends both at Caltech and elsewhere. Many thanks to Emile Oshima, Ubamanyu, Cai Tong Ng, Keree McGuire, Conor Martin, Meredith Xu, Weizhi Yu, Zhilin Zhang, Lucy Cheng, and Corey Zhou. This is by no means a comprehensive list, and there are many more who have touched my life in various ways.

I am especially grateful to the Caltech Chamber Music program and Caltech Orchestra for nourishing my life outside of research. I am deeply indebted to many teachers, including Martin Chalifour, Maia Jasper White, Glenn Price, Robert Ward, and the late Allen Robert Gross. I have fond memories with many friends at Chamber Music and Orchestra, and I would like to thank Miles Chan, Vincent Lee, Sam Davis, Paulina Salazar, Yufeng Du, Bob Gutzman, and Mary Carter, among others, for the opportunities to create wonderful music with them.

Last but not least, I would like to thank my parents, Yumei Jin and Yingfa Zheng, and my brother, Zhangrui Zheng. You are always a safe harbor to me that sustains me through the currents of life.

ABSTRACT

Low-rank matrix recovery problems are prevalent in modern data science, machine learning, and artificial intelligence, and the low-rank property of matrices is widely exploited to extract the hidden low-complexity structure in massive datasets. Compared with Burer-Monteiro factorization in the Euclidean space, using the low-rank matrix manifold has its unique advantages, as it eliminates duplicated spurious points and reduces the polynomial order of the objective function. Yet a few fundamental questions have remained unanswered until recently. We highlight two problems here in particular, which are the global geometry of the manifold and the global convergence guarantee.

As for the global geometry, we point out that there exist some spurious critical points on the boundary of the low-rank matrix manifold $\mathcal{M}_r$, which have rank smaller than r but can serve as limit points of iterative sequences in the manifold $\mathcal{M}_r$. For the least squares loss function, the spurious critical points are rank-deficient matrices that capture part of the eigen spaces of the ground truth. Unlike classical strict saddle points, their Riemannian gradient is singular and their Riemannian Hessian is unbounded.

We show that randomly initialized Riemannian gradient descent almost surely escapes some of the spurious critical points. To prove this result, we first establish the asymptotic escape of classical strict saddle sets consisting of non-isolated strict critical submanifolds on Riemannian manifolds. We then use a dynamical low-rank approximation to parameterize the manifold $\mathcal{M}_r$ and map the spurious critical points to strict critical submanifolds in the classical sense in the parameterized domain, which leads to the desired result. Our result is the first to partially overcome the nonclosedness of the low-rank matrix manifold without altering the vanilla gradient descent algorithm. Numerical experiments are provided to support our theoretical findings.

As for the global convergence guarantee, we point out that earlier approaches to many of the low-rank recovery problems only imply a geometric convergence rate toward a second-order stationary point. This is in contrast to the numerical evidence, which suggests a nearly linear convergence rate start-

ing from a global random initialization. To establish the nearly linear convergence guarantee, we propose a unified framework for a class of low-rank matrix recovery problems including matrix sensing, matrix completion, and phase retrieval. All of them can be considered as random sensing problems of low-rank matrices with a linear measurement operator from some random ensembles. These problems share similar population loss functions that are either least squares or its variant.

We show that under some assumptions, for the population loss function, the Riemannian gradient descent starting from a random initialization with high probability converges to the ground truth in a nearly linear convergence rate, i.e., it takes $O(\log \frac{1}{\epsilon} + \log n)$ iterations to reach an ϵ-accurate solution. The key to establishing a nearly optimal convergence guarantee is closely intertwined with the analysis of the spurious critical points $\mathcal{S}_{\#}$ on $\mathcal{M}_r$. Outside the local neighborhoods of spurious critical points, we use the fundamental convergence tool by the Łojasiewicz inequality to derive a linear convergence rate. In the spurious regions in the neighborhood of spurious critical points, the Riemannian gradient becomes degenerate and the Łojasiewicz inequality could fail. By tracking the dynamics of the trajectory in three stages, we are able to show that with high probability, Riemannian gradient descent escapes the spurious regions in a small number of steps.

After addressing the two problems of global geometry and global convergence guarantee, we use two applications to demonstrate the broad applicability of our analytical tools. The first is the robust principal component analysis problem on the manifold $\mathcal{M}_r$ with the Riemannian subgradient method. The second application is the convergence rate analysis of the Sobolev gradient descent method for the nonlinear Gross-Pitaevskii eigenvalue problem on the infinite dimensional sphere manifold. These two examples demonstrate that the analysis of manifold first-order algorithms can be extended beyond the previous framework, to nonsmooth functions and subgradient methods, and to infinite dimensional Hilbert manifolds. This exemplifies that the insights gained and tools developed for the low-rank matrix manifold $\mathcal{M}_r$ can be extended to broader scientific and technological fields.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. Analysis of asymptotic escape of strict saddle sets in manifold optimization. *SIAM Journal on Mathematics of Data Science*, 2(3):840–871, 2020. doi: 10.1137/19M129437X.
Z. Zhang participated in the conception of the project, performed the analysis, conducted the numerical experiments, and participated in the writing of the manuscript.

[2] Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. Fast global convergence for low-rank matrix recovery via Riemannian gradient descent with random initialization. *arXiv preprint arXiv:2012.15467*, 2020.
Z. Zhang participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[3] Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. Asymptotic escape of spurious critical points on the low-rank matrix manifold. *arXiv preprint arXiv:2107.09207*, 2021.
Z. Zhang participated in the conception of the project, performed the analysis, conducted the numerical experiments, and participated in the writing of the manuscript.

[4] Ziyun Zhang. Exponential convergence of Sobolev gradient descent for a class of nonlinear eigenproblems. *Communications in Mathematical Sciences*, 20(2):377–403, 2022. doi: 10.4310/CMS.2022.v20.n2.a4.
Z. Zhang is the sole author of the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Chapter 1

INTRODUCTION

Low-rank matrix recovery problems are prevalent in modern data science, machine learning, and artificial intelligence. The low-rank property of matrices is widely exploited to extract the hidden low-complexity structure in massive datasets from machine learning, signal processing, imaging science, advanced statistics, information theory, and quantum mechanics. Some excellent overviews of the low-rank recovery problems can be found in the survey papers [39, 68, 80, 116]. Below are a few specific examples.

Matrix sensing. The problem of matrix sensing is one of the canonical model problems for low-rank matrix recovery [39] and has its background in linear matrix equations [30, 116]. The problem can be formulated as follows. Suppose the unknown ground truth matrix is $X \in \mathbb{F}^{n_1 \times n_2}$, where $\mathbb{F}$ can be either $\mathbb{R}$ or $\mathbb{C}$, and X is a low-rank matrix satisfying $\operatorname{rank}(X) = r \ll \min\{n_1, n_2\}$. We know the measurement matrices $\{A_j\}_{j=1}^m \subset \mathbb{F}^{n_1 \times n_2}$, whose entries are drawn i.i.d. from a standard normal distribution, i.e., $\mathcal{N}(0,1)$, if $\mathbb{F} = \mathbb{R}$; or $\frac{\sqrt{2}}{2}\mathcal{N}(0,1) + i\frac{\sqrt{2}}{2}\mathcal{N}(0,1)$, if $\mathbb{F} = \mathbb{C}$. Usually we require that m is at least $O(\min\{n_1, n_2\})$. Let $y = \frac{1}{\sqrt{m}}(\langle A_1, X\rangle, \ldots, \langle A_m, X\rangle)^\top$. The goal is to recover X from y and $\{A_j\}_{j=1}^m$.

Matrix completion. The problem of matrix completion has a long history since [31, 32]. It is still widely used in collaborative filtering for recommendation systems [18], where the goal is to predict users' preferences for unseen items given their reactions to the items they have already seen. In mathematical terms, the problem can be formulated as follows. Suppose $X \in \mathbb{R}^{n_1 \times n_2}$ is the unknown ground truth matrix, and we assume $\operatorname{rank}(X) = r \ll \min\{n_1, n_2\}$. The information we have is the value of X at a small proportion of the entries, whose indices are denoted as Ω. Usually we assume that the index set Ω is generated by a uniform sampling of indices and $|\Omega| \ll n_1 n_2$. The goal is to recover X from the values of Ω only.

Phase retrieval. The Fourier phase retrieval problem originated from optics [56, 81], where it is used in diffraction imaging. More recently, Gaussian phase retrieval has been studied in [34, 122]. The problem can be formulated as a low-rank recovery problem as follows. Suppose $x \in \mathbb{C}^n$ is the unknown ground truth vector. The measurement vectors are $\{a_j\}_{j=1}^m$, whose entries are i.i.d. Gaussian. We only have the magnitude information of the measurements $\{|\langle a_j, x\rangle|\}_{j=1}^m$, but *not* the phase. Denote $A_j = a_j a_j^*$ and $X = xx^*$, then the problem can be seen as recovering a rank-1 matrix X from $(\langle A_1, X\rangle, \dots, \langle A_m, X\rangle)^\top$, which is a low-rank recovery problem.

Robust principal component analysis. This problem is used in image inpainting [113] and video processing [35] and has long been studied in low-rank matrix recovery [33]. We are given a matrix $M \in \mathbb{F}^{n_1 \times n_2}$ which is the sum of a rank-r ground truth L_* and a sparse noise S_*. The goal is to recover L_* and S_* from $M = L_* + S_*$.

Neural networks. Because of the rapid development of machine learning in recent years, it is impossible to identify the single most important application of low-rank recovery in neural networks. Here we just look at one example from the fine-tuning of large language models and generative art models. In a recent work named LoRA [79], it is proposed that large models can be fine-tuned with low-rank updates. Suppose W is the weight matrix consisting of tunable parameters, the goal is to find the best ΔW with $\operatorname{rank}(\Delta W) = r$ such that $W' = W + \Delta W$ optimizes the objective of the neural network. Empirical study [4] shows that the intrinsic dimensionality of large models is actually low, implying that a low-rank approximation of the weight matrices could recover most of the capabilities of the model. LoRA makes the training of large models significantly cheaper and is now widely used in generative models.

The **low-rank matrix manifold** [70, 71] is a useful tool for solving low-rank recovery problems like those above. It has gained popularity in recent years since it gives a neat description of low-rank matrices. Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, and consider the set of all matrices of size $n_1 \times n_2$ with the same rank $r \leq \min\{n_1, n_2\}$. Then this set is a Riemannian manifold, which is a classical result, see for example [91, Example 8.14]. We call it the low-rank matrix manifold[1] and denote it as

$$\mathcal{M}_r = \left\{X \in \mathbb{F}^{n_1 \times n_2} : \operatorname{rank}(X) = r\right\},$$

[1] In some literature it is called the *fixed rank matrix manifold.*

where the subscript r denotes the rank of the matrices in the manifold, and the dimension of the matrices is usually evident from the context. Sometimes we are only interested in the set of $n \times n$ symmetric positive semi-definite (SPSD) matrices $\mathbb{S}_n^+$ or Hermitian positive semi-definite (HPSD) matrices $\mathbb{H}_n^+$, and we can define the corresponding low-rank SPSD/HPSD matrix manifold as

$$\mathcal{M}_r = \left\{ X \in \mathbb{S}_n^+ \text{ or } \mathbb{H}_n^+ : \operatorname{rank}(X) = r \right\}.$$

See Chapter 2, Section 2.1 for more details.

As a Riemannian manifold, $\mathcal{M}_r$ is nonconvex and is locally isomorphic to the Euclidean space. Thus, many nonconvex optimization techniques can be transferred to $\mathcal{M}_r$ without much difficulty. Among them, the most common one is the Riemannian gradient descent, which is the manifold version of the vanilla gradient descent in the Euclidean space. For a function $f : \mathcal{M}_r \to \mathbb{R}$, starting from an initial guess $Z_0 \in \mathcal{M}_r$, the Riemannian gradient descent (RGD) generates a sequence of matrices $\{Z_k\}_{k=0}^{\infty} \subset \mathcal{M}_r$ as follows:

$$Z_{k+1} = R(Z_k - \alpha_k \cdot \operatorname{grad} f(Z_k)).$$

Here $\operatorname{grad} f$ is the Riemannian gradient of the function f on $\mathcal{M}_r$, and α_k is the kth step size. For more details see Section 2.2. This algorithm demonstrates nearly optimal convergence rate and practical flexibility in a number of problems, see e.g., [26, 39, 118, 125, 127] and our work [77].

Compared with the Burer-Monteiro factorization method (the parameterization $X = UV^*$, $U \in \mathbb{F}^{n_1 \times r}$, $V \in \mathbb{F}^{n_2 \times r}$, or $X = UU^*$, $U \in \mathbb{F}^{n \times r}$) in the Euclidean space, using the low-rank matrix manifold has a few advantages. One obvious advantage is that it avoids the duplication of spurious points in the Burer-Monteiro factorization. For example, if $Z_* = UV^*$ is a saddle point or local minima, then any (UQ, VQ) with a unitary $Q \in \mathbb{O}^{r \times r}$ is also a saddle point or local minima in the factorized space, but there is no such problem on the manifold $\mathcal{M}_r$ because we are looking at Z_* itself directly.

Another advantage is that the function of Z is a lower-order polynomial compared with the function of U or (U, V). For example, with the least squares loss function, $f(Z)$ is a quadratic polynomial while $f(U)$ is a quartic (4th order) polynomial. When using the Łojasiewicz inequality (Chapter 5, Section 5.3) to analyze the convergence rate, it is easier to prove a linear convergence rate for quadratic functions.

In addition to the above immediate advantages, a perhaps more remote aspect is the potential extension from $\mathcal{M}_r$ to other Riemannian manifolds. Extracting low-complexity features from high-complexity models is a recurring theme in applied mathematics. Manifold learning [124] has long become a standard technique in modern machine learning. We hope that the insights gained and tools developed for the low-rank matrix manifold $\mathcal{M}_r$ can be extended to other manifolds in broader scientific and technological fields.

The past few years have seen great progress in the study of the low-rank matrix manifold, yet a few fundamental questions have remained unanswered until recently. In particular, we highlight two problems in this thesis, which are the **global geometry** of the manifold and the **global convergence guarantee**.

The problem with **global geometry** lies in the fact that the set $\mathcal{M}_r = \{X \in \mathbb{F}^{n_1 \times n_2} : \operatorname{rank}(X) = r\}$ or $\mathcal{M}_r = \{X \in \mathbb{S}_n^+ \text{ or } \mathbb{H}_n^+ : \operatorname{rank}(X) = r\}$ is not a *closed* set. As we will see in Chapter 2, Section 2.1, the boundary of $\mathcal{M}_r$ consists of matrices with rank less than r. Closedness of the domain is useful in nonconvex optimization because when using an iterative algorithm to optimize a function on the manifold, closedness is necessary to guarantee that the limit point of the iterative sequence is still in the manifold. Indeed, because $\mathcal{M}_r$ is not closed, examples (like Example 2.3.2) demonstrate that some matrices in $\mathcal{M}_s$ with rank $s < r$ can be the limit points of the iterative sequences in $\mathcal{M}_r$. We call them the *spurious critical points*. When the objective function is the least squares function $f(Z) = \frac{1}{2}\|Z - X\|_F^2$, the set of spurious critical points is denoted as $\mathcal{S}_\#$ and can be characterized as follows:

$$\begin{aligned}\mathcal{S}_\# := \{Z_* : \ & Z_* = U_1 D_1 V_1^*, \text{ where } U_1,\ V_1 \text{ and } D_1 \text{ are submatrices of } U_X,\ V_X \text{ and} \\ & D_X \text{ satisfying } U_X = (U_1, U_2),\ V_X = (V_1, V_2),\ D_X = \operatorname{diag}\{D_1, D_2\},\ Z_* \neq X\}.\end{aligned}$$

Here $X = U_X \Sigma_X V_X^*$ is the singular value decomposition (SVD) of X. Section 2.3 gives a more detailed description of these spurious critical points.

Our work [76–78] is among the first to give a thorough analysis of the spurious critical points and their influence on the Riemannian gradient descent algorithm. We discover that the Riemannian gradient becomes singular at the spurious critical points or in the spurious regions in their local neighborhoods. We show that Riemannian gradient descent almost surely escapes some of the spurious critical points. Moreover, with high probability, Riemannian gradient descent avoids these spurious critical points in nearly linear time.

The problem with **global convergence rate** lies in the fact that earlier analysis of the aforementioned low-rank recovery problems only implies a geometric convergence rate toward a second-order stationary point. This is in contrast to the numerical evidence, which suggests a nearly linear convergence rate starting from a global random initialization, see e.g., Figure 1.1.

In our work [77], we show that the key to establishing a nearly optimal convergence guarantee is in fact closely intertwined with the analysis of the spurious critical points $\mathcal{S}_\#$ on $\mathcal{M}_r$. The spurious regions are where the Riemannian gradient becomes degenerate and the Łojasiewicz inequality might fail. Careful treatment is needed to bound the probability that the randomly initialized gradient descent is attracted to the spurious critical points and the number of steps it takes to avoid the spurious regions. We show that using the Riemannian gradient descent to minimize the population loss function for a class of low-rank recovery problems on the manifold $\mathcal{M}_r$, it takes $O(\log n + \log \frac{1}{\epsilon})$ time to converge to an ϵ-accurate solution.

In the following, we discuss each of the results in this thesis in more detail.

1.1 Asymptotic escape of strict saddle sets in manifold optimization

In nonconvex optimization, it is a classical result that first-order algorithms like gradient descent escape from strict saddle points and converge to local minima. This holds true not only in the Euclidean space [89, 90, 112], but also on Riemannian manifolds [41, 123]. Combined with the fact that many low-rank recovery problems actually do not have spurious local minima apart from the ground truth solution [55, 64, 98], this has facilitated the global analysis of first-order algorithms for many applications.

Before our work [76], however, previous analysis on strict saddles mostly focuses on isolated saddle points [89, 90]. There is a lack of explicit treatment for non-isolated continuous saddle sets, despite their prevalence. This motivates us to develop a more general analytic tool for such case.

In Chapter 3, we present a systematic analysis for the asymptotic escape of non-isolated and possibly continuous saddle sets with complicated geometry. We prove that Riemannian gradient descent is able to escape strict critical submanifolds under certain conditions.

An important implication of this saddle analysis is that it paves the way for the asymptotic escape of spurious critical points in Chapter 4. The spurious critical points are different from the classical strict saddle points in that the Riemannian gradient is *singular* around a spurious critical point $Z_* \in \mathcal{S}_\#$. But with a parameterization of the manifold $\mathcal{M}_r$ and a rescaling of the gradient flow, a spurious critical point can be mapped to a classical strict saddle set in the parameterized manifold. In this way, the saddle set analysis serves as an intermediate tool that tackles a much less understood question about the low-rank matrix manifold.

1.2 Asymptotic escape of spurious critical points

Next, we look at the global geometry of the low-rank matrix manifold $\mathcal{M}_r$ and tackle the fundamental problem of nonclosedness of $\mathcal{M}_r$. As we have mentioned before, there exist some matrices with rank smaller than r that can be the limit points of the iterative sequences in $\mathcal{M}_r$. In the SPSD/HPSD case, for the least squares loss function $f(Z) = \frac{1}{2}\|Z - X\|_F^2$, the set of spurious critical points is $\mathcal{S}_\# = \cup_{s=0}^{r-1}\mathcal{S}_s$, where each $\mathcal{S}_s$ can be characterized as

$$\mathcal{S}_s = \big\{Z_\# : \ Z_\# = U_1 D_1 U_1^*, \ U_1 \in \mathbb{F}^{n\times s}, \ D_1 \in \mathbb{F}^{s\times s}\big\}.$$

Here $X = U_X D_X U_X^*$ is an eigenvalue decomposition of X, $U_X = (U_1, U_2)$, $U_1 \in \mathbb{F}^{n\times s}$, $U_2 \in \mathbb{F}^{n\times(r-s)}$ is a block decomposition of U_X, and $D_X = \mathrm{diag}\{D_1, D_2\}$, $D_1 \in \mathbb{R}^{s\times s}$, $D_2 \in \mathbb{R}^{(r-s)\times(r-s)}$ is a block decomposition of D_X.

To see why each matrix in $\mathcal{S}_\#$ can be a limit point of an iterative sequence, one can construct examples explicitly. In Example 2.3.2, using the RGD algorithm $Z_{k+1} = R(Z_k - \alpha \cdot P_{T_{Z_k}}(Z_k - X))$ to minimize the least squares loss function $f(Z) = \frac{1}{2}\|Z - X\|_{\mathrm{F}}^2$ on the manifold $\mathcal{M}_2 = \{Z : Z \in \mathbb{S}_3, \ \mathrm{rank}(Z) = 2\}$, where $X = \mathrm{diag}\{2,\ 1,\ 0\}$, starting from $Z_0 = \mathrm{diag}\{2,\ 0,\ 1\}$, the sequence $\{Z_k\}_{k=0}^{\infty}$ converges to $Z_\# = \mathrm{diag}\{2,\ 0,\ 0\}$, rather than X.

However, with a slightly perturbed initialization, the RGD alogrithm almost always escapes $Z_\#$ and converges to X instead. In fact, when the initial point

Z_0 is sampled on $\mathcal{M}_r$ according to some general random sampling scheme, we observe that convergence to spurious critical points almost never happens. This motivates us to conjecture that their basins of attraction actually have zero measure, and gradient descent almost surely avoids them.

In Chapter 4, we give a partial confirmatory answer to the above conjecture. Our results on the asymptotic escape of spurious critical points can be summarized as follows. We show that when minimizing the least squares loss function, the Riemannian gradient flow and the Riemannian gradient descent with varying stepsize asymptotically escape the rank-(r-1) spurious critical points on the rank-r SPSD or HPSD manifold. More specifically, we have the following informal theorem:

Theorem *(Informal version of Theorems 4.2.1 and 4.3.1)*. The gradient flow Z_t of the least squares loss function and the gradient descent sequence $\{Z_k\}_{k=0}^{\infty}$ with step size $\alpha_k \sim \sigma_r(Z_k)$ almost surely escapes the rank-(r-1) spurious critical points, i.e., Prob $(\lim_{t\to\infty} Z_t \in \mathcal{S}_{r-1}) = 0$ and Prob $(\lim_{k\to\infty} Z_k \in \mathcal{S}_{r-1}) = 0$.

Many previous attempts have been made to deal with the rank-deficient critical points, also called apocalyptic points in some literature[2]. One way is to replace $\mathcal{M}_r$ with $\overline{\mathcal{M}_r}$ and replace tangent space projection with tangent cone projection [118]. Another direction is to modify the algorithm to help it converge to stationary points, for example by using a smooth lift [92] or using numerical rank information [109].

Empirically, however, the original Riemannian gradient descent seems to be robust enough. Even though the spurious critical points can be the limit points of the Riemannian gradient descent algorithm, the required initialization is so special that it is almost impossible under random initialization. In Example 2.3.2 on $\mathcal{M}_2$, from a slightly perturbed initial point with arbitrarily small perturbation, $\{Z_k\}_{k=0}^{\infty}$ converges to X instead of the spurious $Z_\#$.

To understand this phenomenon, it helps to compare it with the asymptotic escape of strict saddle points by gradient descent [89]. The two are remark-

[2] In [92], the set of apocalyptic points on $\overline{\mathcal{M}_r}$ is $\cup_{s<r}\mathcal{M}_s$. We remark that the set of spurious critical points here in Definition 4.1.1 is a subset the apocalyptic points described in [92], i.e., for each s, $\mathcal{S}_s \subsetneq \mathcal{M}_s$. This is because each element in $\mathcal{S}_s$ is not only rank-s, but matches s out of r of X's eigen components. If X has distinct singular values, then $\mathcal{S}_s$ is a finite set.

ably similar, except that the spurious critical points in our context are *not* strict saddle points. Instead, the spurious critical points are singular points where the Riemannian Hessian gets unbounded in certain directions.

To deal with the singularity of the spurious critical points, we propose using the dynamical low-rank approximation [85] to parameterize the gradient flow on the low-rank matrix manifold. We then introduce a rescaled gradient flow to remove the singularity of the ODE system. After rescaling, each spurious critical point becomes a strict critical submanifold in the parameterized domain. Classical saddle escape theorems can then be applied to derive the desired result.

1.3 Fast global convergence for low-rank matrix recovery

The central result of this thesis is the global convergence guarantee for a class of low-rank matrix recovery problems under a unified framework. In Chapter 5, We show that for the population least squares loss function, under certain assumptions, with high probability the Riemannian gradient descent on the manifold $\mathcal{M}_r$ starting from a global random initialization converges to the ground truth in a nearly linear convergence rate, i.e., it takes $O\left(\log \frac{1}{\epsilon} + \mathrm{poly}(\log n)\right)$ iterations to reach an ϵ-accurate solution.

We propose a unified framework for a class of low-rank recovery problems, expressed as the following optimization problem over the low-rank matrix manifold $\mathcal{M}_r$:

$$\min_{Z \in \mathcal{M}_r} f(Z) = \frac{1}{2}\|T(Z) - y\|_2^2, \tag{1.1}$$

where $T : \mathcal{M}_r \to \mathbb{R}^m$ is a linear operator, $T(Z) = \frac{1}{\sqrt{m}}(\langle A_1, Z\rangle, \ldots, \langle A_m, Z\rangle)^\top$, and $y \in \mathbb{R}^m$ with $y_j = \frac{1}{\sqrt{m}}\langle A_j, X\rangle$. The formulation is general and covers many different low-rank matrix recovery problems that we have discussed. Here are a few examples.

(1) Matrix sensing: $T : \mathcal{M}_r \to \mathbb{R}^m$, where $\{A_j\}_{j=1}^m \subset \mathbb{F}^{n_1 \times n_2}$ have entries drawn i.i.d. from $\mathcal{N}(0,1)$, if $\mathbb{F} = \mathbb{R}$; or $\frac{\sqrt{2}}{2}\mathcal{N}(0,1) + i\frac{\sqrt{2}}{2}\mathcal{N}(0,1)$, if $\mathbb{F} = \mathbb{C}$.

(2) Matrix completion: $T : \mathcal{M}_r \to \mathbb{R}^m$, where $\{A_j\}_{j=1}^m \subset \mathbb{F}^{n_1 \times n_2}$ are generated by a uniform sampling of indices $\Omega \subset [n_1] \times [n_2]$ of an $n_1 \times n_2$ matrix. The matrix A_j is the indicator matrix of the j-th sampled entry, which has value 1 at the sampled index and 0 at other indices.

(3) Gaussian phase retrieval: $T : \mathcal{M}_1 \to \mathbb{R}^m$, where $\mathcal{M}_1$ is the symmetric rank-1 matrix manifold, and $\{A_j\}_{j=1}^m \subset \mathbb{F}^{n\times n}$ are rank-1 matrices. In the real case, $A_j = a_j a_j^\top$, where $a_j \in \mathbb{R}^n$ and their entries are drawn i.i.d. from $\mathcal{N}(0,1)$; in the complex case, $A_j = a_j a_j^*$, where $a_j \in \mathbb{C}^n$ and their entries are drawn i.i.d. from $\frac{\sqrt{2}}{2}\mathcal{N}(0,1) + i\frac{\sqrt{2}}{2}\mathcal{N}(0,1)$.

All the above examples can be considered as random sensing problems of low-rank matrices, where T is a linear operator and $\{A_j\}_{j=1}^m$ are drawn from some random distribution. Despite the difference in problem settings and the distributions of $\{A_j\}_{j=1}^m$, their population loss functions share some common properties. Thus we focus on the global convergence behavior of the population loss functions. More specifically, the population loss of matrix sensing and matrix completion is $\mathbb{E}f(Z) = c\|Z - X\|_F^2$ with some positive constant $c > 0$, while the population loss of the phase retrieval problem is $\mathbb{E}f(Z) = c\|Z - X\|_F^2 + \frac{1}{2}(\|Z\|_F - \|X\|_F)^2$ with $c = 1$ or $\frac{1}{2}$ (see Theorem 5.7.2).

$$\min_{Z\in\mathcal{M}_r} F_1(Z) = \frac{1}{2}\|Z - X\|_F^2, \tag{1.2}$$

$$\text{or } \min_{Z\in\mathcal{M}_r} F_2(Z) = \frac{1}{2}(\|Z\|_F - \|X\|_F)^2 + c\|Z - X\|_F^2. \tag{1.3}$$

We are interested in the population loss functions F_1 or F_2 because they can be seen as model problems for the finite sample loss functions. The behavior of the algorithms on the population loss functions to some extent predicts the behavior on the finite-sample loss functions, which can be seen from subsequent chapters. A similar strategy is also employed in [38] and [47]. In Chapter 5, we mostly focus on analyzing Problem (1.2), as it lays the foundation for other population losses. We then briefly discuss Problem (1.3).

Let the sequence $\{Z_k\}_{k=0}^\infty$ be generated by randomly initialized[3] Riemannian gradient descent algorithm:

$$Z_{k+1} = \mathcal{R}\left(Z_k - \alpha_k P_{T_{Z_k}}(\nabla f(Z_k))\right),$$

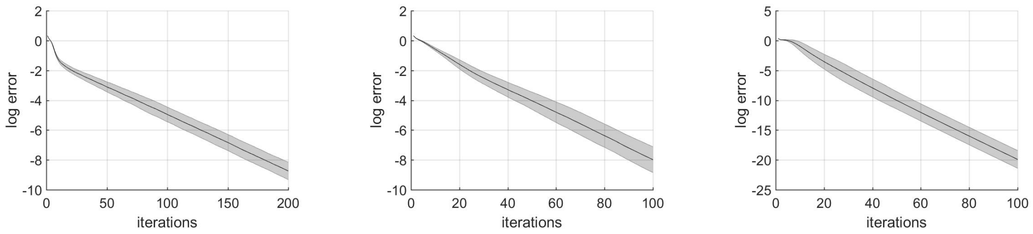
where $P_{T_{Z_k}}$ is the projection onto the tangent space of $\mathcal{M}$ at point Z_k, α_k is the k-th stepsize, and $\mathcal{R} : T_Z \to \mathcal{M}$ is a retraction operator. Below is the main result.

[3] Randomly initialized means the initialization is drawn from the general random distribution (defined in Definition 5.5.1).

Theorem *(Informal version of Theorem 5.1.3). For the population loss* (1.2)*, with high probability no less than* $1 - \frac{1}{poly(n)}$*, the sequence generated by a randomly initialized Riemannian gradient descent needs* $O(poly(\log n) + \log \frac{1}{\epsilon})$ *iterations to reach an* ϵ*-accurate solution of* X*, i.e., to reach* $\|Z - X\|_F \leq \epsilon \|X\|_F$*.*

The above result provides a partial explanation for the mechanism behind the nearly linear convergence rate of vanilla first-order methods and reveals the shared mechanism behind such low-rank matrix recovery problems. The $O(\log n)$ or $O(\text{poly}(\log n))$ term in the number of iterations is mainly because these problems have spurious critical points, and the sequence needs this many iterations to escape these spurious points and their local regions. Our results imply that the randomly initialized Riemannian first-order scheme with high probability converges to the second-order stationary points at a nearly linear rate that is essentially independent of the dimensionality.

Figure 1.1 gives some numerical results obtained using the randomly initialized Riemannian gradient descent to minimize the least squares loss function (1.1) for the three problems. We observe nearly linear convergence in all three experiments. This is consistent with the theoretical results stated above. Our contribution lies in bridging the gap between theory and practice by analyzing the mathematical mechanism behind the fast convergence rate.



(a) *Least squares function* with $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$, n=200, r=10.

(b) *Matrix sensing* with $f(Z) = \frac{1}{2}\|T(Z) - y\|_2^2$, n=100, r=5, m=2500.

(c) *Phase retrieval* with $f(Z) = \frac{1}{2}\|T(Z) - y\|_2^2$, n=100, m=1000.

Figure 1.1: The randomly initialized Riemannian gradient descent. Each log-error band stands for the results from 100 independent experiments.

To further highlight the unique challenges that we overcome, in the following, we review some related works in the literature by topic, and discuss their connection and comparison with our work.

Burer-Monteiro factorization. Earlier approaches to the low-rank matrix recovery problems use convex relaxation [30–32, 34, 116]. Later ones shift focus to nonconvex methods due to their lighter computational cost, and these nonconvex methods mainly rely on the Burer-Monteiro factorization (the parameterization $X = UU^*$, $U \in \mathbb{F}^{n\times r}$ in the Hermitian case) and the global landscape analysis of the corresponding nonconvex objective function [11, 20, 55, 63–65, 98, 122]. In some applications, it has been shown that the landscape does not have spurious local minima [55, 64, 65, 98]. This property, combined with the convergence guarantee of nonconvex optimization [63, 82], leads to a convergence guarantee for second-order stationary points.

However, the convergence rate results obtained by those nonconvex methods are largely sub-linear, in sharp contrast to numerical observations indicating faster, nearly linear convergence. Instead of Burer-Monteiro factorization, we look at the low-rank matrices as a whole on $\mathcal{M}_r$. As we have discussed at the beginning of this chapter, using the manifold $\mathcal{M}_r$ instead of the Burer-Monteiro factorization eliminates duplicated spurious points and reduces the polynomial order of the objective function. Eventually, it enables us to obtain a better convergence guarantee.

Rank-1 versus rank-r. Quite a few previous works explore the asymptotic landscape and exact convergence rate for rank-1 problems, see e.g., [38, 133]. Our results differ from these previous works in that we provide a unified framework of analysis that applies to the general rank-r problem (Theorem 5.1.3) from the low-rank matrix manifold perspective. It is important to note that the general rank-r case is much more challenging than the rank-1 problem. To see the main technical challenge for general rank-r problems, note that while for the rank-1 problem, the core matrix is an 1×1 matrix (a scalar), for the general rank-r problem, it becomes an $r \times r$ matrix. The closed form solutions of some quantities (e.g., the angles between the column spaces) are no longer available, which adds considerable difficulty to the convergence analysis.

Convergence rate. Earlier landscape analysis on the low-rank matrix recovery [20, 55, 64, 65, 98, 122], combined with the convergence guarantee for the nonconvex optimization [63, 82], only implies a geometric convergence rate toward a second-order stationary point. As we have seen in Figure 1.1, numerical evidences suggest a linear convergence rate, and it is common among many low-rank recovery problems. Recent work [38] proves nearly linear convergence rate for the rank-1 phase retrieval problem. This result is consistent with ours, but we extend beyond rank-1 to the general rank-r case. We intend to explore the common mechanism for the linear convergence rate behind many low-rank recovery problems by studying them under a shared framework.

The power of randomness. First-order schemes are widely used in large-scale computation due to their light computational cost. Under certain assumptions, first-order schemes are shown to have fast local convergence to the neighborhood of stationary points [106, 107]. A common problem with first-order schemes is that undesirable critical points, including saddle points and local maxima, could occur. Due to the lack of geometry information around critical points, first-order schemes with bad initialization could get trapped around the undesirable critical points instead of converging to the local minima. However, when augmented with randomness, the first-order schemes work well and have some provable guarantees. Below we discuss two ways of incorporating randomness: the randomly perturbed first-order schemes and the randomly initialized first-order schemes.

1. *Perturbed first-order schemes.* Convergence of perturbed first-order schemes toward second-order stationary points has been studied both in the Euclidean and the Riemannian settings, see [41, 51, 63, 82, 123]. These results show that the general global convergence rate is polynomial and almost dimension-free. Whereas perturbations help perturbed schemes escape the saddles better than non-perturbed first-order schemes in the worst case [51], the perturbations also prevent a very accurate approximation of the ground truth without further (and sometimes complicated) modification.

2. *Randomly initialized first-order schemes.* Though it has been proved that randomly initialized gradient descent asymptotically escapes saddles and only converges to the local minima [76, 89, 90, 112], its convergence rate is much less clear. In the worst case, when the initialization is close to the stable manifold of saddle points, the convergence toward the local minima slows down substantially. Indeed, the authors of the previous work [51] show that, in the worst case, the randomly initialized gradient descent can take exponential time to escape from the saddles. Despite such worst case

scenario, the optimal efficiency of saddle escape behavior in a more general sense remains unclear. A recent answer to this question is given by [38], where it is shown that for the rank-1 phase retrieval problem, gradient descent with random initialization has a nearly linear and almost dimension-free convergence rate, improving upon the previous algebraic convergence rate. This motivates us to investigate the underlying mechanism and establish a similar fast convergence rate for general rank-r matrix recovery problems.

We point out that an alternative way to avoid saddle points is to use Hessian-based schemes and second-order geometric information around the critical points. However, the computational costs of such methods are much higher, and the implementation can be complicated.

Apocalypse-free algorithms. Recent works [92, 109] study the escape of spurious critical points (called apocalyptic points in those works) from a different perspective. They describe the *apocalypse* event where the sequence of iterative points is in $\mathcal{M}_r$ but the limit point has rank less than r and is not stationary. Along this line, it is proposed that a second-order algorithm using a smooth lift [92] or a numerical rank reduction step [109] could avoid spurious critical points and guarantee convergence to the stationary points. We remark that both approaches require major modifications to the first-order algorithm, while we focus on establishing a guarantee for the Riemannian gradient descent algorithm without modification.

1.4 Robust principal component analysis on the manifold

In Chapter 6, we discuss an application of the low-rank matrix manifold $\mathcal{M}_r$ that is beyond the framework of linear measurements and smooth loss functions. We study the Robust Principal Component Analysis (RPCA) problem, whose goal is to recover an unknown low-rank matrix L_* and an unknown sparse matrix S_* from their sum M:

$$\text{Find } L_* \in \mathcal{M}_r,\ S_* \text{ sparse, such that } M = L_* + S_*.$$

The problem of RPCA is interesting for a couple of reasons. One reason comes from our choice of the objective function to promote sparsity. We minimize the (vectorized) l_1 norm as our loss function:

$$f(L) = \|M - L\|_{l_1}, \qquad \text{where } \|X\|_{l_1} := \sum_{i,j} |X_{ij}|.$$

The l_1 norm is a nonsmooth function, but it has a well-defined subgradient (the generalized sign function), which enables us to define its Riemannian subgradient. We thus use the Riemannian subgradient descent algorithm to minimize $f(L)$:

$$Z_{k+1} = R(Z_k - \alpha_k P_{T_{Z_k}}(\partial f(Z_k))),$$

where $P_{T_{Z_k}}(\partial f(\cdot))$ is the Riemannian subgradient of f. We develop a tool for the convergence guarantee of the Riemannian subgradient method on the manifold $\mathcal{M}_r$, which is based on the sharpness and the weak convexity conditions [45]. We show that the Riemannian subgradient method for the l_1 loss function satisfies the sharpness and weak convexity conditions in the local neighborhood of the ground truth L_*. Combined with a spectral initialization that is guaranteed to be in the local neighborhood, we are able to establish the linear convergence rate of the Riemannian subgradient algorithm.

Another reason why RPCA is worth our attention is that the incoherence property of the low-rank matrix L_* is heavily involved in this problem. Assuming the SVD of L_* is $L_* = U\Sigma V^*$, the incoherence parameter is defined as the positive constant μ satisfying

$$\|U\|_{2,\infty} := \max_{1\le i\le m} \|U(i,:)\|_2 \le \frac{\mu\sqrt{r}}{\sqrt{m}}, \qquad \|V\|_{2,\infty} := \max_{1\le i\le n} \|V(i,:)\|_2 \le \frac{\mu\sqrt{r}}{\sqrt{n}}.$$

This gives us the opportunity to take a deep dive into the incoherence of the tangent space of $\mathcal{M}_r$ near L_* and gain some interesting insights. We use concentration tools to show that most elements in the tangent space at an incoherent matrix on $\mathcal{M}_r$ are also incoherent. We also use incoherence to show that when $r = 1$, L_* is guaranteed to be the local minimizer of $f(L) = \|M - L\|_{l_1}$ on the manifold $\mathcal{M}_1$.

1.5 Sobolev gradient descent for a class of nonlinear eigenproblems

In Chapter 7, we discuss another application that further extends beyond the low-rank matrix manifold, as it is posed on an infinite dimensional Hilbert manifold. We explore how the insights gained and tools developed for the low-rank matrix manifold $\mathcal{M}_r$ can be extended to other manifolds in broader scientific and technological fields. Specifically, we demonstrate that the Łojasiewicz inequality tool can be used to derive the convergence guarantee for the manifold gradient descent method here.

The problem of interest is the Gross-Pitaevskii eigenproblem, a well-known example of the nonlinear Schrödinger eigenproblem, which seeks $\lambda \in \mathbb{R}$ and $v \in H_0^1(\Omega)$ that satisfy Equation (7.1):

$$-\Delta v + Vv + \beta |v|^2 v = \lambda v \quad \text{on } \Omega \subset \mathbb{R}^d,$$

where Ω is a bounded region in $\mathbb{R}^d$, $V(x) \geq 0$ is an external trapping potential, and $\beta \geq 0$ is a parameter describing the repulsive interaction between particles. To find the ground state v is equivalent to solving minimization problem (7.2):

$$\min_{\|u\|_{L^2}=1,\ u\in H_0^1(\Omega)} E(u) := \int_\Omega \left(|\nabla u|^2 + V|u|^2 + \frac{\beta}{2}|u|^4 \right) \mathrm{d}x.$$

The constraint set $\{u \in H_0^1(\Omega) : \|u\|_{L^2} = 1\}$ is the unit sphere in $H_0^1(\Omega)$. It can be seen as an infinite dimensional Hilbert manifold. Thus many manifold optimization methods on the Riemannian manifold are readily applicable to this problem, with diverse techniques and rich theories.

In Chapter 7, we use a manifold gradient descent method named the *Sobolev projected gradient descent (Sobolev PGD)* first proposed in [72]:

$$u_{n+1} = R\left((1-\tau_n)\, u_n + \tau_n \cdot \frac{(u_n, u_n)_{L^2}}{(\mathcal{G}_{u_n} u_n, \mathcal{G}_{u_n} u_n)_{a_{u_n}}} \mathcal{G}_{u_n} u_n \right),$$

where R is the retraction onto the manifold, τ_n is the n-th step size, $(\cdot,\cdot)_{a_{u_n}}$ is an adaptive inner product in the tangent space of $\mathcal{M}$, and $\mathcal{G}_{u_n}$ is the Greens operator associated with $(\cdot,\cdot)_{a_{u_n}}$. Their definitions can be found in Section 7.3. The main result is as follows.

Theorem *(Informal version of Theorem 7.3.14).* If initialized with a positive initial guess u_0, the a_u-Sobolev gradient descent which is given by (7.3) converges to the ground state of the eigenproblem (7.1) exponentially fast.

To prove the convergence rate, we introduce the Łojasiewicz inequality tool to this problem. The Łojasiewicz inequality has been discussed in Section 1.3 where it serves as a fundamental convergence tool for low-rank recovery on the low-rank matrix manifold. We rewrite it in a slightly different form in Section 7.2. Using the Łojasiewicz inequality tool, we reveal that the key to exponential convergence is the quadratic nature of the objective energy functional. In other words, regarded as a polynomial, the objective functional should behave like a degree-2 polynomial under the given manifold metric.

The Łojasiewicz inequality tool also makes the Sobolev gradient descent easily applicable to general optimization of high-degree objective or eigenvalue problems other than the Gross-Pitaevskii eigenvalue problem. Its interesting property of making a high-degree polynomial behave like a quadratic one is not specific to a certain problem, but is general. Examples include the biharmonic Schrödinger, the nonlinear Schrödinger with a different order or extra interaction terms, and potentially some general manifold optimization problems. We analyze an example of nonlinear Schrödinger eigenproblem from [17] at the end of the chapter. Extension to more general manifold optimization problems would be of interest in future research.

1.6 Organization of the thesis

The remainder of this thesis is organized as follows.

In Chapter 2, we give a self-contained overview of the low-rank matrix manifold $\mathcal{M}_r$. We introduce the basic properties of the manifold and the basic operations on it. We discuss two Riemannian first-order algorithms on the manifold, i.e., the Riemannian gradient descent and the Riemannian subgradient descent. We give a detailed description of the spurious critical points on the boundary of the manifold. We also include some auxiliary lemmas at the end.

The next few chapters each discuss the results of one section in the introduction. Specifically:

In Chapter 3, we discuss the asymptotic escape of non-isolated strict saddle sets in manifold optimization.

In Chapter 4, we discuss the global geometry of the low-rank matrix manifold $\mathcal{M}_r$, and show the asymptotic escape of Riemannian gradient descent from some of the spurious critical points for the SPSD/HPSD low-rank matrix manifold.

In Chapter 5, we discuss the global convergence guarantee of Riemannian gradient descent on the low-rank matrix manifold $\mathcal{M}_r$. We establish the nearly optimal global convergence rate of RGD for the population loss of a class of low-rank recovery problems. We show that starting from a general random initialization, Riemannian gradient descent converges to the ground truth in a nearly linear convergence rate.

In Chapter 6, we discuss solving the robust principal component analysis problem on the manifold $\mathcal{M}_r$ with the Riemannian subgradient method.

In Chapter 7, we discuss the convergence rate of the Sobolev gradient descent method for the nonlinear Gross-Pitaevskii eigenvalue problem on the infinite dimensional sphere manifold.

Chapter 2

LOW-RANK MATRIX MANIFOLD: AN OVERVIEW

The purpose of this chapter is to give a self-contained overview of the low-rank matrix manifold and related topics. In Section 2.1, we present the definition of the manifold and some basic constructions on the manifold. In Section 2.2, we discuss the Riemannian first-order algorithms that we use. In Section 2.3, we introduce the spurious critical points on the manifold, an important concept that will show up in both asymptotic escape and convergence rate analysis in upcoming chapters.

2.1 Low-rank matrix manifold

The low-rank matrix manifold is defined as $\mathcal{M}_r := \{X \in \mathbb{F}^{m\times n} : \operatorname{rank}(X) = r\}$, where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, and $r \in \{0, \ldots, \min\{m, n\}\}$. Sometimes we consider the manifold of real symmetric positive semi-definite (SPSD) or Hermitian positive semi-definite (HPSD) matrices, which can be defined as $\mathcal{M}_r := \{X \in \mathbb{S}_n^+ \text{ or } \mathbb{H}_n^+ : \operatorname{rank}(X) = r\}$, where $\mathbb{S}_n^+$ or $\mathbb{H}_n^+$ denote the set of n by n SPSD or HPSD matrices respectively.[1] The manifold is slightly different with different $\mathbb{F}$ and with or without symmetry. We use the same notation $\mathcal{M}_r$ throughout the thesis, but clarify its meaning in each chapter. Below we summarize the basic properties of the manifold in each case.

Lemma 2.1.1 (Real, asymmetric case). *Let $\mathcal{M}_r = \{Z \in \mathbb{R}^{m\times n} : rank(Z) = r\}$. Let $\overline{\mathcal{M}_r}$ be its closure. The following is true about $\mathcal{M}_r$ and $\overline{\mathcal{M}_r}$:*

(1) $\overline{\mathcal{M}_r} = \{Z \in \mathbb{R}^{m\times n} : rank(Z) \le r\}$;

(2) $\mathcal{M}_r$ is dense in $\overline{\mathcal{M}_r}$;

(3) $\mathcal{M}_r$ is connected;

(4) The local dimension of $\mathcal{M}_r$ is $(m + n - r)r$;

(5) The boundary of $\mathcal{M}_r$ is $\overline{\mathcal{M}_r} \setminus \mathcal{M}_r = \cup_{0\le s<r}\mathcal{M}_s$.

[1] In some literature, the notation is $\mathcal{S}_+(r, n)$ for the SPSD fixed rank manifold, and $\mathcal{H}_+(r, n)$ for the HPSD fixed rank manifold. see e.g., [105].

Proof.

(2) For each $Z \in \overline{\mathcal{M}_r} \setminus \mathcal{M}_r$, it can be approached by a sequence of rank-r matrices $\{Z_k\} \subset \mathcal{M}_r$ such that $\lim_{k\to\infty} Z_k = Z$.

(3) Consider

$$\begin{aligned} \widetilde{\Phi_r} : \mathbf{SO}(m, \mathbb{R}) \times \mathbb{R}_+^r \times \mathbf{SO}(n, \mathbb{R}) &\to \mathbb{R}^{m\times n} \\ (\widetilde{U}, \boldsymbol{\sigma}_r, \widetilde{V}) &\mapsto \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top, \end{aligned}$$

where $\mathbf{SO}(m, \mathbb{R})$ is the real orthogonal group in dimension m, $X = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top$ is the full singular value decomposition of X, $\boldsymbol{\sigma}_r \in \mathbb{R}^r$, $\boldsymbol{\sigma}_r(i) \neq 0$, $i = 1, \cdots, r$, and

$$\widetilde{\Sigma} = \begin{pmatrix} \mathrm{diag}(\sigma_r) & \\ & 0_{(m-r)\times(n-r)} \end{pmatrix}.$$

Since $\mathbf{SO}(m, \mathbb{R}) \times \mathbb{R}_+^r \times \mathbf{SO}(n, \mathbb{R})$ is connected and $\widetilde{\Phi_r}$ is continuous, its orbit $\mathcal{M}_r$ is connected.

(4) Consider the compact singular value decomposition $X = U\mathrm{diag}(\boldsymbol{\sigma}_r)V^\top$, where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, and $\boldsymbol{\sigma}_r$ is in descending order. The local dimension is $r + \frac{(2m-r-1)r}{2} + \frac{(2n-r-1)r}{2} = (m+n-r)r$.

(5) It is obviously true.

□

Lemma 2.1.2 (Complex, non-Hermitian case). *Let $\mathcal{M}_r = \{Z \in \mathbb{C}^{m\times n} : rank(Z) = r\}$, and let $\overline{\mathcal{M}_r}$ be its closure. The following is true about $\mathcal{M}_r$ and $\overline{\mathcal{M}_r}$:*

(1) $\overline{\mathcal{M}_r} = \{Z \in \mathbb{C}^{m\times n} : rank(Z) \leq r\}$;

(2) $\mathcal{M}_r$ is dense in $\overline{\mathcal{M}_r}$;

(3) $\mathcal{M}_r$ is connected;

(4) The local dimension of $\mathcal{M}_r$ is $(2m + 2n - r)r$;

(5) The boundary of $\mathcal{M}_r$ is $\overline{\mathcal{M}_r} \setminus \mathcal{M}_r = \cup_{0\leq s<r}\mathcal{M}_s$.

Proof.

(4) Consider the compact singular value decomposition $X = U\mathrm{diag}(\boldsymbol{\sigma}_r)V^*$. The local dimension is $r + \frac{(4m-r-1)r}{2} + \frac{(4n-r-1)r}{2} = (2m + 2n - r)r$.

□

Lemma 2.1.3 (Real symmetric case). *Let $\mathcal{M}_r = \{Z \in \mathbb{S}_n : rank(Z) = r\}$ and let $\overline{\mathcal{M}_r}$ be its closure. Then*

(1) $\overline{\mathcal{M}_r} = \{Z \in \mathbb{S}_n : rank(Z) \leq r\}$;

(2) $\mathcal{M}_r$ is dense in $\overline{\mathcal{M}_r}$;

(3) $\mathcal{M}_r$ has $r + 1$ disjoint branches and each branch is connected;

(4) The local dimension of $\mathcal{M}_r$ is $\frac{(2n-r+1)r}{2}$;

(5) The boundary of $\mathcal{M}_r$ is $\overline{\mathcal{M}_r} \setminus \mathcal{M}_r = \cup_{0\leq s<r}\mathcal{M}_s$.

Proof.

(3) Consider the set of matrices that has p positive eigenvalues and q negative eigenvalues, $p + q = r$. Define

$$\begin{aligned}\Psi_{p,q} : \mathbf{GL}^+(n, \mathbb{R}) &\longrightarrow \mathbb{S}^n \\ P &\mapsto PI_{p,q}P^*\end{aligned}$$

where $\mathbf{GL}^+(n, \mathbb{R})$ is the real positive-determinant group in dimension n, and

$$I_r = \begin{pmatrix} I_p & & \\ & -I_q & \\ & & 0_{(n-r)\times(n-r)} \end{pmatrix}.$$

Thus, the orbit of each $\Psi_{p,q}$ is connected. The tuple (p, q) is called the *signature* of the matrix. However, matrices with different signatures are not path-connected on $\mathcal{M}_r$ (they are path-connected only on $\overline{\mathcal{M}_r}$). So $\mathcal{M}_r$ has $r + 1$ branches corresponding to the orbits of $\Psi_{r,0}, \Psi_{r-1,1}, \cdots, \Psi_{0,r}$.

(4) Consider $X \in \mathcal{M}_r$ and let $X = UDU^\top$ be its compact eigenvalue decomposition with descending eigenvalues. Consider the mapping

$$\Phi_s : (U, \boldsymbol{\sigma}_r) \mapsto UDU^*.$$

The local dimension is $r + \frac{(2n-r-1)r}{2} = \frac{(2n-r+1)r}{2}$.

□

Lemma 2.1.4 (Complex Hermitian case)**.** *Let $\mathcal{M}_r = \{Z \in \mathbb{H}_n : rank(Z) = r\}$ and let $\overline{\mathcal{M}_r}$ be its closure. Then*

(1) $\overline{\mathcal{M}_r} = \{Z \in \mathbb{H}_n : rank(Z) \leq r\}$;

(2) $\mathcal{M}_r$ is dense in $\overline{\mathcal{M}_r}$;

(3) $\mathcal{M}_r$ has r + 1 disjoint branches and each branch is connected;

(4) The local dimension of $\mathcal{M}_r$ is $\frac{(4n-r+1)r}{2}$;

(5) The boundary of $\mathcal{M}_r$ is $\overline{\mathcal{M}_r} \setminus \mathcal{M}_r = \cup_{0 \leq s < r} \mathcal{M}_s$.

Proof.

(4) Consider $X \in \mathcal{M}_r$ and let $X = UDU^*$ be its compact eigenvalue decomposition with descending eigenvalues. Consider the mapping

$$\Phi_s : (U, \boldsymbol{\sigma}_r) \mapsto UDU^*.$$

The local dimension is $r + \frac{(4m-r-1)r}{2} = \frac{(4m-r+1)r}{2}$.

□

The tangent space of $\mathcal{M}_r$ is defined as follows, see [125] for further references.

Definition 2.1.5 (Tangent space, non-Hermitian case)**.** Let $X \in \mathcal{M}_r$, $X = U\Sigma V^\top$ (or $X = U\Sigma V^*$). Let $\mathcal{U} = \mathrm{Col}(U)$, $\mathcal{V} = \mathrm{Col}(V)$ be the column spaces of U and V respectively. Then the *tangent space* of $\mathcal{M}_r$ at X is

$$T_X\mathcal{M}_r = (\mathcal{U} \otimes \mathcal{V}) \oplus (\mathcal{U} \otimes \mathcal{V}^\perp) \oplus (\mathcal{U}^\perp \otimes \mathcal{V}).$$

We use the abbreviation T_X instead of $T_X\mathcal{M}_r$ when the manifold $\mathcal{M}_r$ is clear from context. The projection operator onto the tangent space can be characterized as:

$$P_{T_X}(Y) = P_U \cdot Y + Y \cdot P_V - P_U \cdot Y \cdot P_V,$$

where $P_U = UU^*$ and $P_V = VV^*$ are the projections onto the subspaces $\mathcal{U}$ and $\mathcal{V}$ respectively.

Definition 2.1.6 (Tangent space, Hermitian case)**.** Let $X \in \mathcal{M}_r$, $X = UDU^\top$ (or $X = UDU^*$), $\mathcal{U} = \mathrm{Col}(U)$. Then the *tangent space* of $\mathcal{M}_r$ at X is

$$T_X\mathcal{M}_r = (\mathcal{U} \otimes \mathcal{U}) \oplus (\mathcal{U} \otimes \mathcal{U}^\perp) \oplus (\mathcal{U}^\perp \otimes \mathcal{U}).$$

The projection operator onto the tangent space can be characterized as

$$P_{T_X}(Y) = P_U \cdot Y + Y \cdot P_U - P_U \cdot Y \cdot P_U,$$

where $P_U = UU^*$ is the projection onto the subspace $\mathcal{U}$.

We are sometimes interested in the closure $\overline{\mathcal{M}_r}$ instead of $\mathcal{M}_r$ itself. A few reasons might come into play as to why people might favor $\overline{\mathcal{M}_r}$ over $\mathcal{M}_r$. First, $\mathcal{M}_r$ itself is not a closed set. Iterative algorithms for nonconvex optimization generate a sequence that converges to a solution, but closedness is necessary for the sequence to have a limit. Second, a sequence might simply cross the boundary of $\mathcal{M}_r$ at a certain iterate and the rank might fall below r. For these reasons, $\overline{\mathcal{M}_r}$ are sometimes used. Since $\overline{\mathcal{M}_r}$ is constructed by "gluing together" all lower-rank matrix manifolds, it needs some special treatment at $\mathcal{M}_s$ $(s < r)$ in order to make up for the deficient dimension. In addition to the classical tangent space, we can define *tangent cone* at these lower-dimensional sets. A reference on this is [118].

Definition 2.1.7 (Tangent cone, non-Hermitian case)**.** Let $X \in \mathcal{M}_s \subset \overline{\mathcal{M}_r}$ where $s < r$, $X = U\Sigma V^\top$ (or $X = U\Sigma V^*$), $\mathcal{U} = \mathrm{Col}(U)$, $\mathcal{V} = \mathrm{Col}(V)$. Then the *tangent cone* of $\overline{M_r}$ at X is

$$T_X\overline{\mathcal{M}_r} = T_X\mathcal{M}_s \oplus \{\eta : \eta \in \mathcal{U}^\perp \otimes \mathcal{V}^\perp,\ \mathrm{rank}(\eta) = r - s\}.$$

The projection onto the tangent cone is the projection onto the tangent space plus a rank (r-s) principal component, i.e.

$$P_{T_X\overline{\mathcal{M}_r}}(Y) = P_{T_X\mathcal{M}_s}(Y) + Y_{r-s},$$

where Y_{r-s} is a best rank (r-s) approximation of $Y - P_{T_X\mathcal{M}_s}(Y)$ in the Frobenious norm.

Definition 2.1.8 (Tangent cone, Hermitian case)**.** Let $X \in \mathcal{M}_s \subset (\overline{\mathcal{M}_r})$ where $s < r$, $X = UDU^\top$ (or $X = UDU^*$), $\mathcal{U} = \mathrm{Col}(U)$. Then the *tangent cone* of $\overline{M_r}$ at X is

$$T_X\overline{M_r} = T_X\mathcal{M}_s \oplus \{\eta : \eta \in \mathcal{U}^\perp \otimes \mathcal{U}^\perp,\ \mathrm{rank}(\eta) = r - s\}.$$

The projection onto the tangent cone is

$$P_{T_X\overline{\mathcal{M}_r}}(Y) = P_{T_X\mathcal{M}_s}(Y) + Y_{r-s},$$

where Y_{r-s} is a best rank (r-s) approximation of $Y - P_{T_X\mathcal{M}_s}(Y)$ in the Frobenious norm.

When $\overline{\mathcal{M}_r}$ is involved, we use P_{T_z} for both $P_{T_z\mathcal{M}_r}$ and $P_{T_z\overline{\mathcal{M}_r}}$ as long as there is no confusion.

Next, we define the retraction operation. It is also called the projection onto the manifold in some literature. The retraction operation is not unique, but we usually require that it satisfies the following first-order retraction property.

Definition 2.1.9 (First-order retraction)**.** Let $\mathcal{M}$ be an arbitrary Riemannian manifold and let $\|\cdot\|$ be the norm of the ambient Banach space of $\mathcal{M}$. Let T_Z be the tangent space (or tangent cone) of $\mathcal{M}$ at Z. We say that $\mathcal{R}$ satisfies the *first-order retraction property*, if for any $Z \in \mathcal{M}$, $\xi \in T_Z$,

$$\lim_{\alpha\to 0^+} \frac{\|\mathcal{R}(Z + \alpha\xi) - (Z + \alpha\xi)\|}{\alpha} = 0. \tag{2.1}$$

The above definition applies to any manifold $\mathcal{M}$ and not just $\mathcal{M}_r$. When it comes to the low-rank matrix manifold $\mathcal{M}_r$, a natural retraction on $\mathcal{M}_r$ is the following best rank-r approximation under the Frobenius norm:

$$\mathcal{R}_N(Z + \xi) = \arg\min_{Y\in\mathcal{M}_r} \|Z + \xi - Y\|_F.$$

We use $\mathcal{R}_N$ as our $\mathcal{R}$ for $\mathcal{M}_r$ in the rest of the thesis. It not only satisfies the first-order retraction property, but is actually second order, see also [110, 118, 125]:

$$\mathcal{R}(Z + \alpha\xi) = Z + \alpha\xi + \alpha^2\eta + O(\alpha^3).$$

Vandereycken [125] provides an explicit second-order approximation $\mathcal{R}_N^{(2)}$ to this second-order retraction $\mathcal{R}(\cdot)$ and shows that $\mathcal{R}(Z+\xi) = \mathcal{R}_N^{(2)}(Z+\xi) + O(\|\xi\|^3)$.

To define and analyze Riemannian nonconvex optimization methods on the manifold $\mathcal{M}_r$, we need to define the Riemannian gradient and Hessian. They are different from their Euclidean counterparts because Riemannian manifolds are only locally isomorphic to an Euclidean space, and careful treatment is needed to find out what corresponds to the Euclidean gradients on the manifold. We first introduce the concepts in their most general form, then look at the special case that is actually used in practical computation.

Definition 2.1.10 (Levi-Civita connection)**.** The *Levi-Civita connection* $\widetilde{\nabla}_\xi \eta$, acting on two vectors or vector fields η, ξ in the tangent bundle $T\mathcal{M}$ of a Riemannian manifold $\mathcal{M}$, is the unique affine connection on $\mathcal{M}$ that preserves the metric and is torsion-free.

Note that the notation $\widetilde{\nabla}$ is not to be confused with ∇, which we use to denote the gradient in the ambient space.

Definition 2.1.11 (Riemannian gradient)**.** Given $f : \mathcal{M} \to \mathbb{R}$ where $\mathcal{M}$ is an arbitrary Riemannian manifold, the *Riemannian gradient* of f is the vector field $\mathrm{grad} f$, such that for any vector field Y on $\mathcal{M}$,

$$\langle \mathrm{grad} f, Y \rangle = Y(f),$$

where $\langle \cdot, \cdot \rangle$ is the metric on $\mathcal{M}$ and $Y(\cdot)$ is the vector field action, i.e., $Y(f) = \sum_i Y_i \frac{\partial f}{\partial E_i}$ for a basis $\{E_i\}$.

The good news is that Riemannian gradient is equivalent to the tangent space projection of the embedded gradient in the ambient space if they exist, i.e.,

$$\mathrm{grad}\ f(Z) = P_{T_Z}(\nabla f(Z)). \tag{2.2}$$

Furthermore, if the metric of $\mathcal{M}$ is inherited from the ambient space, then the Levi-Civita connection on $\mathcal{M}$ is the tangent space projection of the Levi-Civita connection (natural gradient) of the ambient space. In other words, for $\eta, \xi \in T\mathcal{M}$, we have

$$\widetilde{\nabla}_\xi \eta = P_{T_Z}(\nabla \eta[\xi]).$$

Therefore, if we let $\mathcal{M}_r$ inherit the metric from its ambient Euclidean space, i.e., $\langle A, B\rangle = \mathrm{trace}(A^* B)$ and $\|A\| = \|A\|_F$, then the Levi-Civita connection becomes the Euclidean derivative and the Riemannian gradient is just the projected gradient. Because of this, the Riemannian gradient descent method is sometimes called the projected gradient descent method [76]. Equation (2.2) is how we calculate the Riemannian gradient on $\mathcal{M}_r$ in practice.

Definition 2.1.12 (Riemannian Hessian)**.** Given a function $f : \mathcal{M} \to \mathbb{R}$, the *Riemannian Hessian* of f at point Z is $\mathrm{Hess}\, f(Z) : T_Z\mathcal{M} \to T_Z\mathcal{M}$ defined by

$$\mathrm{Hess}\, f(Z)[\xi] = \widetilde{\nabla}_\xi \mathrm{grad}\, f(Z), \tag{2.3}$$

where $\widetilde{\nabla}_{(\cdot)}(\cdot)$ is the Levi-Civita connection on $\mathcal{M}$.

Proposition 2.1.13. *If the retraction is second-order, i.e.,*

$$P_{T_Z}\left(\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\mathcal{R}_Z(\alpha\xi)\mid_{\alpha=0}\right) = 0,$$

then

$$Hess\ f(Z) = Hess\ (f \circ \mathcal{R}_Z)(0). \tag{2.4}$$

In particular, (2.4) is true for the low-rank matrix manifold, and this can sometimes make the Riemannian Hessian easier to compute, see also [125]. It is proved in [2] that in the case of the low-rank matrix manifold, (2.4) recovers Definition 2.1.12.

2.2 Riemannian first-order algorithms

In this subsection, we discuss two nonconvex optimization algorithms on the low-rank matrix manifold, namely the Riemannian gradient descent method and the Riemannian subgradient descent method.

Riemannian gradient descent

This Riemannian gradient descent (RGD) method is the gradient descent on the Riemannian manifold, and it has long been studied in data science and machine learning problems [24, 118, 125, 127]. Assume we are given a differentiable objective function $f : \mathcal{M} \to \mathbb{R}$ to be minimized, where $\mathcal{M}$ can be a general Riemannian manifold. We start from an initial guess $Z_0 \in \mathcal{M}$. Using the Riemannian gradient descent method, the iterative sequence $\{Z_k\}_{k=0}^{K}$ is generated as follows:

$$Z_{k+1} = \mathcal{R}\left(Z_k - \alpha_k \text{grad} f(Z_k)\right), \tag{2.5}$$

where $\text{grad} f(Z_k)$ is the Riemannian gradient of the function f at $Z_k \in \mathcal{M}$ as defined in Definition 2.1.11, α_k is the k-th stepsize, and $\mathcal{R} : T_Z \to \mathcal{M}$ is a retraction operator.

When the manifold is embedded in an ambient space and its Riemannian metric is inherited from the inner product in the ambient space, by (2.2), the Riemannian gradient descent can be written as follows:

$$Z_{k+1} = \mathcal{R}\left(Z_k - \alpha_k P_{T_{Z_k}}(\nabla f(Z_k))\right). \tag{2.6}$$

Here ∇ is the derivative in the ambient space of the manifold and $P_{T_{Z_k}}$ is the projection onto the tangent space (or tangent cone at a rank-deficient point) of $\mathcal{M}$ at point Z_k.

It is worth mentioning that there exists other manifold optimization techniques. For example, when it comes to $\mathcal{M}_r$, some manifold optimization methods skip the projection step $P_{T_{Z_k}}$ and compute the eigenvalue decomposition directly. This is also called the "hard retraction". We choose the current Riemannian gradient descent algorithm with "soft retraction" mainly due to two reasons:

1) The projected gradient is the true Riemannian gradient, as is evident from (2.2);

2) The projected gradient is also cheaper in computation. Namely, solving $\mathcal{R}(Z_k - \alpha_k \text{grad} f(Z_k))$ involves calculating SVD of a $n_1 \times n_2$ matrix, while $\mathcal{R}(Z_k - \alpha_k P_{T_{Z_k}}(\text{grad} f(Z_k)))$ only involves that of a $2r \times 2r$ matrix, as mentioned in the previous literature [127]. Since $r \ll \min\{n_1, n_2\}$, the soft retraction is of lighter computational cost.

More specifically, to see why RGD has light computational cost, assume that the SVD of Z_k is $Z_k = U_k \Sigma_k V_k^*$ and denote $W_k := \mathrm{grad} f(Z_k) \in \mathbb{F}^{n_1 \times n_2}$, we have

$$\begin{aligned} Z_{k+1} &= \mathcal{R}(Z_k - \alpha P_{T_{Z_k}}(W_k)) \\ &= \mathcal{R}\left(U_k \Sigma_k V_k^* - \alpha(U_k U_k^* W_k + W_k V_k V_k^* - U_k U_k^* W_k V_k V_k^*)\right) \\ &= \mathcal{R}\Big(U_k(\Sigma_k - \alpha U_k^* W_k V_k)V_k^* - \alpha U_k((I - V_k V_k^*)W_k^* U_k)^* \\ &\quad - \alpha((I - U_k U_k^*)W_k V_k)V_k^*\Big) \\ &= \mathcal{R}\left(\begin{pmatrix} U_k & Q_{k,2} \end{pmatrix} \begin{pmatrix} \Sigma_k - \alpha U_k^* W_k V_k & -\alpha R_{k,1}^* \\ -\alpha R_{k,2} & 0 \end{pmatrix} \begin{pmatrix} V_{k,1}^* \\ Q_{k,1}^* \end{pmatrix}\right). \end{aligned}$$

Assume $(I - V_k V_k^*)W_k^* U_k = Q_{k,1} R_{k,1}$ and $(I - U_k U_k^*)W_k V_k = Q_{k,2} R_{k,2}$ are the QR factorizations of the respective matrices. Notice that $Q_{k,2}^* U_k = \mathbf{0}_{r \times r}$, $Q_{k,1}^* V_k = \mathbf{0}_{r \times r}$. Therefore, to compute SVD of $Z_k - \alpha P_{T_{Z_k}}(W_k)$ it only involves solving the SVD of $\begin{pmatrix} \Sigma_k - \alpha U_k^\top W_k V_k & -\alpha R_{k,1}^* \\ -\alpha R_{k,2} & 0 \end{pmatrix}$, which is only a $2r \times 2r$ matrix. The symmetric/Hermitian versions of the algorithm can also be derived similarly by replacing SVD with eigenvalue decomposition.

Riemannian subgradient descent

The Riemannian subgradient method is a variant of the Riemannian gradient descent method for nonsmooth functions with generalized gradient. Assume that $f : \mathcal{M} \to \mathbb{R}$ is a Lipschitz function and has generalized gradient ∂f, and $\mathcal{M}$ inherits its Riemannian metric from the inner product of ambient space, then the Riemannian subgradient method can be written as

$$Z_{k+1} = R(Z_k - \alpha_k P_{T_{Z_k}}(\partial f(Z_k))). \tag{2.7}$$

A common case where the subgradient is needed is when the objective function contains l_1 norm. The l_1 norm is differentiable almost everywhere except when an entry is zero. The generalized gradient of the l_1 norm is the entrywise generalized sign function. An example can be found in Chapter 6, where we solve the Robust Principal Component Analysis (RPCA) problem by minimizing the vectorized l_1 norm using the Riemannian subgradient method on $\mathcal{M}_r$.

2.3 Spurious critical points

One of the important findings in our work is the existence of some singular critical points for the Riemannian first-order algorithms on the manifold

$\mathcal{M}_r$, which we call the *spurious critical points*. Our study of the spurious critical points was motivated by an analysis of the population loss of a class of low-rank recovery problems on the manifold.

We have seen in Section 2.1 that $\mathcal{M}_r$ is *not* a closed set, and the boundary of $\mathcal{M}_r$ consists of matrices with rank smaller than r, which are not in $\mathcal{M}_r$ themselves. In other words, $\overline{\mathcal{M}_r}\backslash\mathcal{M}_r = \cup_{s=0}^{r-1}\mathcal{M}_s \not\subset \mathcal{M}_r$. For this reason, there is no guarantee that the limit point of an iterative sequence will converge to a rank-r ground truth instead of being stuck at some lower-rank spurious points. In particular, when minimizing the least squares loss function $f(Z) = \frac{1}{2}\|Z - X\|_{\mathrm{F}}^2$ with $X \in \mathcal{M}_r$, there exist some points with rank smaller than r which could also serve as the limit points of minimizing sequences. These spurious critical points are lower-rank matrices that captures part of the eigen components of the ground truth. This phenomenon was first reported in [71] and later attracted more research interest.

The following lemma tells us that the critical points of $f(Z) = \frac{1}{2}\|Z - X\|_F^2$ in Riemannian optimization consist of the ground truth X and the set of spurious critical points denoted as $\mathcal{S}_\#$, and each matrix in $\mathcal{S}_\#$ is a lower-rank matrix that captures part of the eigen components of the ground truth.

Lemma 2.3.1 (Spurious critical points)**.** *Consider the Riemannian gradient descent for the function $f(Z) = \frac{1}{2}\|Z - X\|_F^2$ with fixed step size $\alpha_k \equiv \alpha$. Let $X = UDV^*$ be the SVD of $X \in \mathbb{F}^{n_1\times n_2}$, where $D \in \mathbb{R}^{r\times r}$ is a non-singular diagonal matrix (here the diagonals are not necessarily in descending order) and $U \in \mathbb{F}^{n_1\times r}$, $V \in \mathbb{F}^{n_2\times r}$ are orthonormal. Then,*

1) *There are two types of fixed points: one is the ground truth $Z = X$, and the other consists of the set*

   $$\mathcal{S}_\# := \{Z_* : Z_* = U_1D_1V_1^*, \text{ where } U_1,\ V_1 \text{ and } D_1 \text{ are submatrices of } U_X,\ V_X \text{ and } D_X \text{ satisfying } U_X = (U_1, U_2),\ V_X = (U_1, V_2),\ D_X = diag\{D_1, D_2\},\ Z_* \neq X\};$$

2) *Specifically, if X has distinct eigenvalues*[2]*, then $\mathcal{S}_\#$ has cardinality $|\mathcal{S}_\#| = 2^r - 1$. Assume that $X = \sum_{i=1}^r d_iu_iv_i^*$, then $\mathcal{S}_\# = \{Z_* : Z_* = \sum_{i=1}^r d_i\eta_iu_iv_i^*, \text{with } \eta \in \{0,1\}^r \text{ and } \eta \neq (1,\ 1,\ \ldots, 1)^*\}$.*

[2] Which means that the eigenvalues of X all have algebraic multiplicity equal to 1.

In the SPSD/HPSD case, let $X = UDU^*$ be the eigenvalue decomposition of X where $D \in \mathbb{R}^{r\times r}$ is a non-singular diagonal matrix and $U \in \mathbb{F}^{n\times r}$ is orthonormal. Then the set of spurious critical points is

$$\mathcal{S}_{\#} := \big\{Z_* : \; Z_* = U_1 D_1 U_1^*, \text{ where } U = (U_1, U_2), \; D = diag\{D_1, D_2\}, \; Z_* \neq X\big\}.$$

If X has distinct eigenvalues, then $\mathcal{S}_{\#} = \{Z_* \; : \; Z_* = \sum_{i=1}^{r} d_i \eta_i u_i u_i^*, \textit{with } \eta \in \{0,1\}^r$ and $\eta \neq (1,\, 1,\, \ldots, 1)^*\}$.

Proof. We only prove it for the general non-Hermitian case. The proof for the SPSD/HPSD case is similar and we omit the details here.

1) It is obvious that Z is a fixed point of $Z_{k+1} = \mathcal{R}(Z_k - \alpha P_{Z_k}(\text{grad} f(Z_k)))$ if and only if $P_{T_z}(\text{grad} F_1(Z)) = 0$. Denote $Z = U_z \Sigma_z V_z^*$ and $X = UDV^*$, then for any $\xi \in T_Z$, there exists $\Delta_1 \in \mathbb{F}^{n_1\times r}$ and $\Delta_2 \in \mathbb{F}^{n_2\times r}$, such that $\xi = U_z\Delta_1^* + \Delta_2 V_z^*$. Simple calculation gives

$$\begin{aligned}
&\quad P_{T_z}(\text{grad} F_1(Z)) = 0 \Longleftrightarrow \langle P_{T_z}(Z - X), \xi\rangle = 0, \text{ for all } \xi \in T_Z \\
&\Longleftrightarrow \langle Z - X, U_z\Delta_1^* + \Delta_2 V_z^*\rangle = 0, \text{ for all } \xi = U_z\Delta_1^* + \Delta_2 V_z^* \in T_Z \\
&\Longleftrightarrow \text{tr}((V_z\Sigma_z - VDU^*U_z)\Delta_1^* + (\Sigma_z U_z^* - V_z^* VDU^*)\Delta_2), \text{ for all } \Delta_1 \in \mathbb{F}^{n_1\times r}, \Delta_2 \in \mathbb{F}^{n_2\times r} \\
&\Longleftrightarrow V_z\Sigma_z - VDU^*U_z = \Sigma_z U_z^* - V_z^* VDU^* = 0 \\
&\Longleftrightarrow U_z^* X = \Sigma_z V_z^* \text{ and } U_z\Sigma_z = XV_z.
\end{aligned}$$

This implies $P_{U_z}(X) = P_{U_z}(Z) = Z$ and $P_{V_z}(Z^*) = Z^* = P_{V_z}(X^*)$. Assume

$$X = \begin{pmatrix} U_z & \widetilde{U}_z \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \begin{pmatrix} V_z^* \\ \widetilde{V}_z^* \end{pmatrix}.$$

Then we get $X_{11} = \Sigma_z$, $X_{12} = 0$ and $X_{21} = 0$. Therefore, we have

$$Z = U_1 D_1 V_1^*, \text{with } U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}, D = \text{diag}\,\{D_1, D_2\} \text{ and } V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}.$$

2) If X has distinct eigenvalues, then $\mathcal{S}$ consists of the points $Z_* = \sum_{i=1}^{r} d_i \eta_i u_i v_i^*$, where $\eta \in \{0,1\}^r$ and η is not $(1, 1, ..., 1)^* \in \mathbb{R}^r$. So $|\mathcal{S}| = 2^r - 1$.

□

The set $\mathcal{S}_\#$ consists of $2^r - 1$ points (including $Z_* = 0$) if the eigenvalues of X are distinct, or contains some submanifolds of $\mathcal{M}_r$ when at least one eigenvalue has multiplicity more than one. We call them "spurious" critical points because they are not stable, and when the tangent cone is taken into consideration they are not true fixed points. More importantly, it is very unlikely that the sequence generated by the randomly initialized Riemannian gradient descent will converge to any one of these spurious fixed points. Below is an example.

Example 2.3.2. Assume that $n = 3$, $r = 2$. Let

$$X = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad Z_0 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then the $\{Z_k\}_{k=0}^{\infty}$ generated by the Riemannian gradient descent and their limit point are given by

$$Z_k = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & (1-\alpha)^k \end{pmatrix}, \qquad Z_* := \lim_{k\to\infty} Z_k = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

We see that Z_* is a spurious critical point. Note that even though each Z_k is in $\mathcal{M}_2$, their limit is in $\mathcal{M}_1$.

Figure 2.1 is a visualization of the gradient $\|P_{T_Z}(Z - X)\|_F$ in the neighborhood of a spurious Z_*. We can see that the gradient is essentially singular near Z_*. There is only one direction in which the sequence converges to Z_*. Along other directions, the Riemannian gradient remains large and the RGD diverges from Z_*. We point out that such property of these spurious fixed points is very similar to that of *strict saddle points* in many nonconvex optimization problems, although they are not exactly the same because the gradient of the spurious critical point is singular.

Example 2.3.2 demonstrates that we cannot assume that the sequence generated by the Riemannian gradient descent always stays in the interior of $\mathcal{M}_r$ and converges to a minimum. However, numerical evidence shows that the Riemannian gradient descent almost always escapes the spurious critical points $\mathcal{S}_\#$ and converges to the global minimum. In Chapter 4, we will establish the asymptotic escape of Riemannian gradient descent from

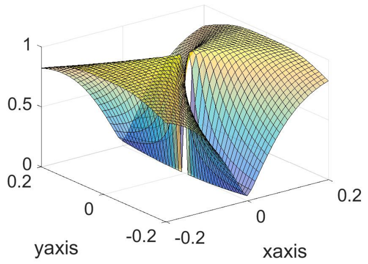


Figure 2.1: Magnitude of the gradient in the neighborhood of a spurious fixed point

some of these spurious critical points; in Chapter 5, we will further establish the nearly linear rate in which Riemannian gradient descent escapes from spurious critical points and converges to the local minima.

2.4 Auxiliary lemmas

Here we list some auxiliary lemmas about matrix analysis that come in handy in future chapters.

Lemma 2.4.1 (The SinΘ Theorem, [44]). *Let A be a Hermitian operator. Assume that*

$$A = \begin{pmatrix} E_0 & E_1 \end{pmatrix} \begin{pmatrix} A_0 & 0 \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} E_0^* \\ E_1^* \end{pmatrix}$$

is an invariant subspace decomposition (i.e., a generalized eigenvalue decomposition) of A. Let

$$B = A + \Delta, \quad B = \begin{pmatrix} F_0 & F_1 \end{pmatrix} \begin{pmatrix} B_0 & 0 \\ 0 & B_1 \end{pmatrix} \begin{pmatrix} F_0^* \\ F_1^* \end{pmatrix}.$$

Let Θ_0 be the angle matrix between subspaces E_0 and F_0. Define the residual as

$$R := BE_0 - E_0 A_0.$$

If there is an interval $[\beta, \alpha]$ and $\delta > 0$, such that the spectrum of A_0 lies entirely in $[\beta, \alpha]$, while that of B_1 lies entirely in $(-\infty, \beta - \delta] \cup [\alpha + \delta, +\infty)$, then for every unitary-invariant norm $\|\cdot\|$, we have

$$\delta \|sin\Theta_0\| \le \|R\|.$$

In particular, this holds true for the matrix 2-norm and the Frobenius norm.

Lemma 2.4.2 (Hoffman-Wielandt Theorem)**.**

1) *Assume $Z, Z' \in \mathbb{R}^{n\times n}$ or $\mathbb{C}^{n\times n}$ are normal matrices, and their corresponding ordered spectra are $\{\lambda_j\}$ and $\{\tilde{\lambda}_j\}$. Then, we have*

$$\sqrt{\sum_{j=1}^{n} |\tilde{\lambda}_j - \lambda_j|^2} \leq \|Z' - Z\|_F.$$

2) *Assume Z, $Z' \in \mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$, and denote their eigenvalues in descending order as $\{\sigma_j\}$ and $\{\tilde{\sigma}_j\}$. Then, we have*

$$\sqrt{\sum_{j=1}^{n} |\tilde{\sigma}_j - \sigma_j|^2} \leq \|Z' - Z\|_F.$$

Chapter 3

ASYMPTOTIC ESCAPE OF STRICT SADDLE SETS IN MANIFOLD OPTIMIZATION

In this chapter, we present some analysis on the asymptotic escape of strict saddle sets in manifold optimization using Riemannian gradient descent. The main contribution here is that we extend previous analysis on the escape of strict saddles to include non-isolated and possibly continuous saddle sets with complicated geometry. We prove that Riemannian gradient descent is able to escape strict critical submanifolds under certain conditions on the geometry and the distribution of the saddle point sets. We also show that the Riemannian gradient descent may fail to escape strict saddles under weaker assumptions even if the saddle point set has zero measure.

As examples of this saddle set analysis, we apply it to the phase retrieval problem on the low-rank matrix manifold, prove that there is only a finite number of saddles, and that in a specific region, they are strict saddles with high probability. We also apply this saddle set analysis to a variational eigenvalue problem on the sphere manifold, which is a special case of the Gross-Pitaevskii eigenvalue problem in Chapter 7.

Beyond the above examples, a more important implication of this saddle analysis is that it paves the way for the asymptotic escape of spurious critical points in Chapter 4. As has been discussed in Section 2.3, the spurious critical points are significantly different from the classical strict saddle points in that the Riemannian gradient is *singular* around a spurious critical point $Z_* \in \mathcal{S}_\#$. It is important to keep in mind that any result on classical strict saddle points cannot be directly applied to the spurious critical points because of their singularity. Interestingly, with a parameterization of the manifold $\mathcal{M}_r$ and a rescaling of the gradient flow, a spurious critical point can be mapped to a strict saddle set in the parameterized manifold. In this way, the saddle set analysis in this chapter acts as an intermediate tool that tackles a much less understood question about the low-rank matrix manifold.

Organization of this chapter. We have given a brief introduction of this problem in Section 1.1. The rest of this chapter is organized as follows. In Section 3.1, we further elaborate on the background of the problem and related works. Section 3.2 contains the result on the asymptotic escape of Riemannian gradient descent from isolated strict saddle points, and Section 3.3 contains the result for non-isolated strict saddle sets and strict critical submanifolds. In Section 3.4, phase retrieval is analyzed as an example of asymptotic escape of saddles on $\overline{\mathcal{M}_r}$. We extend the application to other manifolds and a broader scope of problems in Section 3.5. We make some final discussion in Section 3.6.

3.1 Background

As has been discussed in Section 1.1, in this chapter, we provide a systematic analysis for the asymptotic escape of non-isolated and possibly continuous saddle sets with complicated geometry. This is motivated by the fact that before this work (published in [76]), previous analysis on strict saddles mostly focuses on isolated saddle points, while non-isolated continuous saddle sets have not been thoroughly studied.

Specifically, we prove that Riemannian gradient descent is able to escape strict critical submanifolds under certain conditions. These conditions are concerned with some geometric properties of the saddle set or the distribution of these saddle points. These conditions are necessary to guarantee the asymptotic escape of the strict saddles by the RGD in manifold optimization. However, these conditions are not stringent and are usually satisfied by common applications. We compare our conditions with those of the recent work [112], and point out that these two are consistent. We also give some examples that violate the conditions and result in failures of asymptotic escape, for the purpose of theoretical interest.

What lies at the core of this asymptotic analysis is an interesting interplay of dynamical systems and nonconvex optimization, and a translation of languages from the Morse theory [12, 13, 40] into gradient flows and further into gradient descents. Although these tools were initially developed to study homology, they have provided invaluable insight into the converging/escaping sets of strict saddle points with nontrivial geometry. We draw inspiration from them and propose a new unified tool to analyze asymptotic convergence and escape from saddle points.

We are aware that there is a parallel line of research on stochastic/perturbed gradient descent, as well as Riemannian stochastic gradient descent. A few works including [41, 63, 82, 123] show that the stochastic/perturbed gradient descent is a powerful tool to avoid saddles and does not impose any constraint on the geometry of saddle sets as we do. The reason our analysis focuses on the unperturbed gradient descent is that only by eliminating the perturbation effect can we single out the essential property of gradient descent itself. The development of a thorough asymptotic theory for this vanilla RGD algorithm is crucial toward the understanding of why vanilla RGD works sufficiently well in many applications.

As an application of our asymptotic escape analysis for strict saddles, we consider the phase retrieval problem [56, 66, 81, 119] that has received considerable attention in the recent years. We combine the perspectives of Riemannian manifold optimization [24] and landscape analysis [98, 122] to derive new results. We analyze the saddle points of the phase retrieval problem on the low-rank matrix manifold. Surprisingly, there are only a finite number of saddles and they are all strict saddles with very high probability. Our analysis provides an explanation for the robust performance of the RGD in solving the phase retrieval problem as a low-rank optimization problem, a phenomenon that has been observed in previous applications reported in the literature [98].

Although our primary focus is the low-rank matrix manifold, the asymptotic convergence to the minimum and the escape of strict saddles or strict critical submanifolds are valid on any finite dimensional Riemannian manifold. In particular, the properties of the RGD are well preserved if the manifold is embedded in a Banach space and inherits the Riemannian metric from its ambient space. Common examples of manifold optimization include optimization problems on the sphere, the Stiefel manifold, the Grassmann manifold, the Wasserstein statistical manifold [94], the flag manifold for multiscale analysis [130], and the low-rank tensor manifold [74, 102]. Applications can be found in many fields including physics, statistics, quantum information, and machine learning.

To illustrate this, we consider the eigenvalue problems as minimization problems on the unit sphere and the Stiefel manifold. In the examples, the eigenstates of the linear/nonlinear eigenvalue problems are strict saddle sets on the sphere manifold or Stiefel manifold. This can either be proved analytically or verified numerically. We apply the saddle set escape theorem to these problems. We demonstrate that RGD can act as an acceleration method by formulating a simultaneous eigen-solver on the Stiefel manifold. We observe a considerable speedup in the convergence of the RGD method on the Stiefel manifold.

3.2 Asymptotic escape of isolated saddle points

This section is a self-contained overview of the classical results for the asymptotic escape of strict saddle points by gradient flow and gradient descent. Note that we only intend to include the results for the vanilla gradient descent. We do not cover the perturbed or stochastic gradient descent, as they are less relevant to our problem.

We emphasize that the spurious critical points in Lemma 2.3.1 are *not* classical strict saddle points. It is because the Riemannian gradient at the spurious critical points is singular. Therefore, the theorems and lemmas in this section are *not* directly applicable to the spurious critical points. Nevertheless, these theorems and lemmas will be used in an indirect manner, on a rescaled system where the singularity is removed, see Chapter 4.

Assume we are given a function $f : \mathcal{M} \to \mathbf{R}$ where $\mathcal{M}$ can be a general manifold. We start from a proper initial guess $Z_0 \in \mathcal{M}$. Consider the Riemannian gradient descent (RGD) method with a fixed step size α. Let φ be the iteration operation:

$$Z_{n+1} = \varphi(Z_n) := R(Z_n - \alpha P_{T_{Z_n}}(\nabla f(Z_n))).$$

Here ∇f is the *embedded* gradient of f in its ambient Banach space , $P_{T_{Z_n}}$ is the projection onto the tangent space of $\mathcal{M}$ at point Z_n, α_n is the n-th step size, and $\mathcal{R} : T_Z \to \mathcal{M}$ is a first-order retraction as defined in Definition 2.1.9.

When $\max_n\{\alpha_n\} \to 0$, the limit of the gradient descent is the gradient flow, characterized as

$$\frac{\mathrm{d}}{\mathrm{d}t} Z_t = -\nabla f(Z_t), \qquad Z_t \in \mathcal{M}.$$

We first look at the result on the stable and unstable manifolds of the gradient flow at a hyperbolic point.

Theorem 3.2.1 ([114, The *Center Manifold Theorem*])**.** *Let $\rho \in C^a(E)$ where E is an open subset of $\mathbb{R}^n$ containing the origin and $a \geq 1$. Let $x(t) = \phi_t(x_0)$ be the gradient flow of the system $\dot{x} = \rho(x)$. Suppose that $\rho(0) = 0$ and that $D\rho(0)$ has k eigenvalues with negative real part, j eigenvalues with positive real part, and $m = n - k - j$ eigenvalues with zero real part. Then there exist*

(1) A k-dimensional stable manifold $W^s(0)$ of class C^a tangent to the stable subspace E^s at 0, where for all $x_0 \in W^s(0)$,

$$\lim_{t \to +\infty} \phi_t(x_0) = 0;$$

(2) A j-dimensional unstable manifold $W^u(0)$ of class C^a tangent to the unstable subspace E^u at 0, where for all $x_0 \in W^u(0)$,

$$\lim_{t \to -\infty} \phi_t(x_0) = 0;$$

(3) And an m-dimensional center manifold $W^c(0)$ of class C^a tangent to the center subspace E^c at 0.

Furthermore, $W^c(0)$, $W^s(0)$ and $W^u(0)$ are invariant under the gradient flow.

Next, we introduce the counterpart of the previous results for the gradient descent. The strict saddle point is defined as follows. It basically says that a strict saddle point is a hyperbolic point of the iteration function.

Definition 3.2.2 (Strict saddle point)**.** Consider a function $f : \mathcal{M} \to \mathbb{R}$ defined on a manifold $\mathcal{M}$. We call $Z \in \mathcal{M}$ a *strict saddle point* of f, if

1. $P_{T_Z}(\nabla f(Z)) = 0$;
2. Hess $f(Z)$ has at least one negative eigenvalue.

The result on the Riemannian gradient descent (RGD) and the strict saddles points is as follows.

Theorem 3.2.3 (RGD asymptotically only converges to a local minimum)**.** *Let $f : \mathcal{M} \to \mathbb{R}$ be a C^2 function on $\mathcal{M}$. Let $\{Z_k\}_{k=0}^{\infty}$ be the sequence generated by the Riemannian gradient descent algorithm on $\mathcal{M}$. Suppose that f has either* *finitely many saddle points, or countably many saddle points in a compact submanifold of $\mathcal{M}$, and all saddle points of f are strict saddles as is defined in Definition 3.2.2. Let $\mathcal{A}$ denote the set of strict saddles. Then we have*

$$Prob(\lim_{k\to\infty} Z_k \in \mathcal{A}) = 0.$$

In other words, the RGD with a random initialization almost never converges to a saddle point Z_* as long as Z_* is a strict saddle.

To prove this result, the main tool is the stable manifold theorem on low-rank matrix manifolds, which is an extension of similar results in the Euclidean spaces.

Theorem 3.2.4. *Let $\varphi : \mathcal{M} \to \mathcal{M}$ be a smooth diffeomorphism of a finite dimensional smooth manifold $\mathcal{M}$, and p is a fixed point of φ. Assume that*

$$T_p\mathcal{M} = T_p^s\mathcal{M} \oplus T_p^c\mathcal{M} \oplus T_p^u\mathcal{M}, \tag{3.1}$$

which is the invariant splitting of $T_p\mathcal{M}$ into contracting, centering, and expanding subspaces corresponding to eigenvalues of magnitude less than, equal to, and greater than 1. Let

$$T_p^{cs}\mathcal{M} := T_p^s\mathcal{M} \oplus T_p^c\mathcal{M}.$$

Then we have

$$W_p^s(\varphi) := \{x \in \mathcal{M} | \lim_{n\to\infty} \varphi^n(x) = p\}$$

is an immersed submanifold of $\mathcal{M}$ and $T_pW_p^s(\varphi) \subseteq T_p^{cs}\mathcal{M}$. We call it the (generalized) stable manifold *of p with respect to φ.*

For those who are interested in the proof, a detailed one for the Euclidean case can be found in Theorem III.7 in [120], and the extension to the manifold is similar to [13, Theorem 4.15].

Proof of Theorem 3.2.3. From Theorem 3.2.4, if φ is a diffeomorphism, and $T_{Z_*}\mathcal{M}$ has an invariant splitting as in (3.1) with a nonempty expanding subspace $T_{Z_*}^u\mathcal{M}$, then $W_{Z_*}^s(\varphi)$, the stable set of Z_*, will be a lower dimensional submanifold in $\overline{\mathcal{M}}$. Then, the converging set of Z_* will have measure 0 with respect to the manifold, and any random initialization of RGD will escape such a strict saddle point almost surely.

We now look at the splitting of $T_{Z_*}\mathcal{M}$. The diffeomorphic property of φ is actually contained in the proof of the former.

Given $\xi \in T_{Z_*}\mathcal{M}$, for any $\tilde{\xi}$ that satisfies $Z_* + \tilde{\xi} \in \mathcal{M}$, $P_{T_{Z_*}}(\tilde{\xi}) = \xi$, we have

$$\begin{aligned}\mathrm{Hess}f(Z_*)[\xi] + O(\|\xi\|^2) &= \widetilde{\nabla}_\xi \,\mathrm{grad}f(Z_*)\\ &= P_{T_{Z_*}}(\nabla \mathrm{grad}f(Z_*)[\xi])\\ &= P_{T_{Z_*}}(\mathrm{grad}f(Z_* + \tilde{\xi}) - \mathrm{grad}f(Z_*)) + O(\|\xi\|^2)\\ &= P_{T_{Z_*}}(\mathrm{grad}f(Z_* + \tilde{\xi})) + O(\|\xi\|^2).\end{aligned}$$

Note that $\mathrm{grad}f(Z_*) = 0$ since Z_* is a critical point. Therefore, for $\varphi(Z_n) = R(Z_n - \alpha P_{T_{Z_n}}(\nabla f(Z_n)))$,

$$\begin{aligned}D\varphi_{Z_*}[\xi] &= P_{T_{Z_*}}(\varphi(Z_* + \tilde{\xi}) - \varphi(Z_*)) + \mathrm{o}(\|\xi\|)\\ &= P_{T_{Z_*}}(R(Z_* + \tilde{\xi} - \alpha\ \mathrm{grad}f(Z_* + \tilde{\xi})) - Z_*) + \mathrm{o}(\|\xi\|)\\ &= P_{T_{Z_*}}(Z_* + \tilde{\xi} - \alpha\ \mathrm{grad}f(Z_* + \tilde{\xi}) + \mathrm{o}(\|\xi\|) - Z_*)\\ &= P_{T_{Z_*}}(\tilde{\xi} - \alpha\ \mathrm{grad}f(Z_* + \tilde{\xi})) + \mathrm{o}(\|\xi\|)\\ &= \xi - \alpha\ \mathrm{Hess}f(Z_*)[\xi] + \mathrm{o}(\|\xi\|).\end{aligned}$$

We have

$$D\varphi_{Z_*}[\xi] = \xi - \alpha \mathrm{Hess}(f)(Z_*)[\xi],$$

i.e.

$$D\varphi_{Z_*} = I - \alpha\ \mathrm{Hess}(f)(Z_*).$$

Thus Z_* being strict saddle implies that, by choosing α sufficiently small (but depending on the eigenvalues of $\mathrm{Hess}f(Z_*)$), $D\varphi_{Z_*}$ has at least one expanding subspace coresponding to an eigenvalue whose magnitude is greater than 1.

Now, φ is a diffeomorphism at Z_* because, if we choose α small enough so that $\|\mathrm{Hess}f(Z_*)\| < \frac{1}{\alpha}$, then $D\varphi$ is always invertible and bounded. If there are only finitely many strict saddle points, or there are countably infinite number of them in a compact region, $\|\mathrm{Hess}(f)(Z_*)\|$ shall be upper bounded, and such an α is always attainable.

Using Theorem 3.2.4, the set of points on $\mathcal{M}$ that converge to Z_* is a lower dimensional submanifold in $\mathcal{M}$, which has measure 0. We could safely deduce that, by randomly sampling a start point Z_0 in $\mathcal{M}$, the probability of converging to a strict saddle point is 0, i.e.

$$\mathrm{Prob}(\lim_{k\to\infty} Z_k = Z_*) = 0.$$

Since there are only countably many strict saddle points, $\cup_{Z_*\in S} W^S_{Z_*}(\varphi)$ still has measure 0. So the RGD with a randomly sampled starting point converges to any point in S with probability 0. This proves the first argument.

As for the second argument, since the step size is a constant α, the only stationary points of the algorithm are first-order critical points of the loss function. The local maximizers are ruled out by the descent property of gradient descent. So if the limit point exists, it is a local minimizer with probability 1. □

3.3 Non-isolated strict saddle sets and strict critical submanifolds

As is mentioned in Section 3.1, it is very common that there are more than a countable number of strict saddle points, e.g., when they form a submanifold, or a more complicated set, with Lebesgue measure 0. Empirical evidences show satisfactory convergence of the RGD to its minimum, which indicates successful escape from these strict saddles. But there is a lack of theoretical analysis to confirm this observation. In the following, we will use some further results from the Morse theory and its extensions to provide an analytical tool for this purpose.

Definition 3.3.1 (Critical submanifold)**.** For $f : \mathcal{M} \mapsto \mathbb{R}$, a connected submanifold $\mathcal{N} \subset \mathcal{M}$ is called a *critical submanifold* of f if every point Z in $\mathcal{N}$ is a critical point of f, i.e., $\operatorname{grad} f(Z) = 0$ for any $Z \in \mathcal{N}$.

Definition 3.3.2 (Strict critical submanifold)**.** A critical submanifold $\mathcal{N}$ of f is called a *strict critical submanifold*, if $\forall Z \in \mathcal{N}$,

$$\lambda_{\min}(\mathrm{Hess}\, f(Z)) \le c < 0,$$

where $\lambda_{\min}(\cdot)$ takes the smallest eigenvalue, and $c = c(\mathcal{N})$ is a uniform constant for all $Z \in \mathcal{N}$ depending only on $\mathcal{N}$.

Analogous to the stable/unstable manifolds of critical points in Theorem 3.2.4, we may define stable/unstable manifolds of critical submanifolds.[1]

[1] The reader shall be careful while distinguishing different "manifolds": the domain of the function f is a manifold, and the critical set of f is now a submanifold, but the names of stable/unstable manifolds are given regardless of the domain of f.

Definition 3.3.3 (Generalized stable/unstable manifold)**.** Let $\varphi : \mathcal{M} \to \mathcal{M}$ be a smooth diffeomorphism of $\mathcal{M}$. Then for a submanifold of $\mathcal{N} \subset \mathcal{M}$, the *stable manifold* and *unstable manifold* of $\mathcal{N}$ with respect to φ are defined as

$$W^s_{\mathcal{N}}(\varphi) := \{x \in \mathcal{M} | \lim_{n\to\infty} \varphi^n(x) \in \mathcal{N}\},$$
$$W^u_{\mathcal{N}}(\varphi) := \{x \in \mathcal{M} | \lim_{n\to-\infty} \varphi^n(x) \in \mathcal{N}\}.$$

Given a nontrivial strict critical submanifold $\mathcal{N}$ of f, at every point $p \in \mathcal{N}$, the tangent space is split as

$$T_p\mathcal{M} = T_p\mathcal{N} \oplus \nu_p\mathcal{N},$$

where $\nu_p\mathcal{N}$ is the normal space of $\mathcal{N}$ at p immersed in $\mathcal{M}$. Similar to Equation (3.1), it is further split into

$$T_p\mathcal{M} = T_p\mathcal{N} \oplus (\nu^s_p\mathcal{M} \oplus \nu^c_p\mathcal{M} \oplus \nu^u_p\mathcal{M}).$$

To arrive at a result similar to that stated in Theorem 3.2.4, it suffices to define

$$T^{cs}_p\mathcal{M} := T_p\mathcal{N} \oplus (\nu^s_p\mathcal{M} \oplus \nu^c_p\mathcal{M}),$$

and notice that $T_pW^s_{\mathcal{N}}(\varphi) \subseteq T^{cs}_p\mathcal{M}$ for any $p \in \mathcal{N}$. Since $T^{cs}_p\mathcal{M}$ is dimension deficient by the definition of strict critical submanifold, any random initialization still falls into the converging set of $\mathcal{N}$ with probability 0. Because the union of a finite number of 0-measure sets still has measure 0, the above result works well with countably many critical submanifolds. To sum up, we have the following theorem.

Theorem 3.3.4. *Let $f : \mathcal{M} \to \mathbb{R}$ be a C^2 function on $\mathcal{M}$. Let $\{Z_k\}_{k=0}^{\infty}$ denote the sequence generated by the Riemannian gradient algorithm (2.6). Suppose that $f : \mathcal{M} \to \mathbb{R}$ has either finitely many critical submanifolds, or countably many critical submanifolds in a compact region of $\mathcal{M}$, and all of them are strict critical submanifolds as defined in Definition 3.3.3. Let $\mathcal{A}$ denote the union of strict critical submanifolds. Then we have*

$$Prob(\lim_{k\to\infty} Z_k \in \mathcal{A}) = 0.$$

We remark that the results on the asymptotic escape of saddle points in the Euclidean space, e.g., the results in [89], can be seen as special cases of Theorem 3.3.4.

For situations that are even more complicated than those stated in the above theorem, it is conjectured that the transversality relationship of submanifolds can be exploited to find out the succession relationship of critical sets. We refer the reader to Section 3.7 for some useful tools and interesting insights in this direction.

Finally, we point out that the number of critical submanifolds being countable is an essential condition, but not a binding one. Of course, one reason of this statement is that it is often satisfied in practice. Namely, in the known applications with very complicated saddle geometries, e.g. matrix factorization and nonlinear eigenproblems, the saddles can still be grouped into countably many points or submanifolds. In those cases, either Theorem 3.2.3 or Theorem 3.3.4 is applicable.

But even from a purely theoretical point of view, the number of strict critical submanifolds being uncountable is unlikely to happen. This is in accordance with the result of a recent work [112]. The result explicitly includes the case of "uncountably many critical points", but from the viewpoint of submanifolds, such result belongs to the case of "countably many submanifolds" in our Theorem 3.3.4. (A submanifold can contain uncountably many points, but is still a single object to escape.) This can also be inferred from the use of a countable subcover in Theorem 10 of [112] and the subsequent proof of the main theorem, where the convergence set to any saddle is categorized into a countable number of stable manifolds.

To further illustrate this point, here we give some interesting examples. The saddle sets in Example 3.3.5 occupy a zero measure set in the whole manifold. They cannot be assembled into countably many connected submanifolds. We will analyze and see why they cannot be almost surely avoided.

Example 3.3.5. Let $\mathcal{M} = [-1, 2] \times [-1, 1] \subset \mathbb{R}^2$ be a rectangular region, viewed as a 2-dimensional submanifold of $\mathbb{R}^2$. Then the tangent space $T\mathcal{M}$ equals $\mathcal{M}$. To construct the function f on $\mathcal{M}$, we need the 1-dimensional Smith-Volterra-Cantor ("fat Cantor") set V in $[0, 1]$. The construction is as follows:

1) Remove the middle interval of length $\frac{1}{4}$ from $[0, 1]$, and the remaining set is $[0, \frac{3}{8}] \cup [\frac{5}{8}, 1]$;

2) Remove 2 middle subintervals of length $\frac{1}{4^2}$ from the 2 remaining intervals, and the remaining set is $[0, \frac{5}{32}] \cup [\frac{7}{32}, \frac{3}{8}] \cup [\frac{5}{8}, \frac{25}{32}] \cup [\frac{27}{32}, 1]$;

3) Remove 4 middle subintervals of length $\frac{1}{4^3}$ from the 4 remaining intervals;

4) ...

A visualization of the construction is given in Figure 3.1a. The construction is different from that of the classical Cantor set in that we remove proportionally shorter subintervals, instead of subintervals proportional to the mother interval. Therefore, V has positive measure in $\mathbb{R}$, $meas(V) = \frac{1}{2}$, while the classical Cantor set has zero measure. Still, V is nowhere dense.

We look for a synthetic objective function on $\mathcal{M}$ in the form

$$f : \mathcal{M} \to \mathbb{R}, \quad f(x, y) = -p(x) + y^2,$$

where $p(x)$ is a function of certain regularity on the 1D interval $x \in [-1, 2]$. Consider two examples:

(A) Define $p_A(x) = 0$ for $x \in V$. As V is a closed set, write $V^c = [-1, 2] \backslash V = (\bigcup_\alpha (a_\alpha, b_\alpha)) \cup [-1, 0) \cup (1, 2]$ as the disjoint union of intervals. On each interval (a, b), let

$$p_A(x) = \begin{cases} (x-a)^2, & \text{for} \quad a < x \le a + \frac{(b-a)}{4}; \\ C_1(x - \frac{a+b}{2})^4 + C_2(x - \frac{a+b}{2})^2 + C_3, & \text{for} \quad a + \frac{(b-a)}{4} < x \le b - \frac{(b-a)}{4}; \\ (b-x)^2, & \text{for} \quad b - \frac{(b-a)}{4} < x < b, \end{cases}$$

where $C_1 = \frac{8}{(b-a)^2}$, $C_2 = -2$, $C_3 = \frac{5(b-a)^2}{32}$. See a visualization in Figures 3.1b and 3.1c.

(B) Similar to (A), but on each interval (a, b), let

$$p_B(x) = \begin{cases} (x-a)^4, & \text{for} \quad a < x \le a + \frac{(b-a)}{4}; \\ C_1(x - \frac{a+b}{2})^6 + C_2(x - \frac{a+b}{2})^4 + C_3, & \text{for} \quad a + \frac{(b-a)}{4} < x \le b - \frac{(b-a)}{4}; \\ (b-x)^4, & \text{for} \quad b - \frac{(b-a)}{4} < x < b, \end{cases}$$

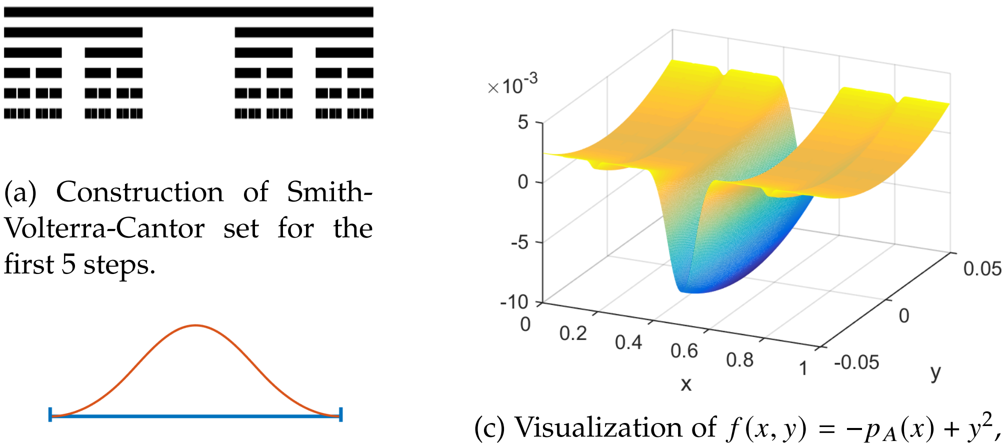
where $C_1 = \frac{512}{3(b-a)^4}$, $C_2 = -\frac{24}{(b-a)^2}$, $C_3 = \frac{11(b-a)^2}{96}$.

It is easy to see that both functions $p_i(x)$, $i = A$ or B, satisfy $p(x) \geq 0$ and $p(x) = 0$ if and only if $x \in V$. Thus for $f_i(x, y) = -p_i(x) + y^2$, the saddle set of f is $S = V \times [0]$. Viewed in the 2-dimensional manifold, it has zero measure. But the converging set of S is $W_S^s(\varphi) = V \times [-1, 1]$. It has positive measure in $\mathcal{M}$: $meas(W_S^s(\varphi)) = 1$. If we start the RGD algorithm with a uniform random initialization, the probability that $\{Z_n\}_{n=0}^\infty$ end up toward a saddle is

$$Pr(\lim_{n\to\infty} Z_n \in S) = \frac{1}{6} > 0.$$

So what happens? The reason that gradient descent fails to escape such a saddle set is well hidden. Specifically, in Example (A), $p_A(x)$ is only C^1 but not C^2. For each $x \in V$, the second derivatives of $p_A(x)$ on two sides are not equal. One side of x is an open interval in V^c, so the second derivative is 2; while on the other side x is the limit point of a sequence $\{x_j\}_{j=1}^\infty \subset V$, and the second derivative is not well-defined. As for Example (B), $p_B(x)$ is C^3 over $[-1, 2]$ and thus $f_B(x, y)$ satisfies the regularity requirements. However, $p''_B(x) = 0 \ \forall x \in V$, so $x \in V$ are not *strict* saddles.

A loosely relevant discussion of the above constructions can be found in [117], Exercise 5.21. This example is purely synthetic, but it raises a healthy warning as to how much the assumptions can be relaxed while the escape from saddle sets is still valid.



(a) Construction of Smith-Volterra-Cantor set for the first 5 steps.

(b) Example of $p_A(x)$ on a single interval.

(c) Visualization of $f(x, y) = -p_A(x) + y^2$, where saddle set is the middle of the "plateaus".

Figure 3.1: Illustration of Example 3.3.5.

3.4 Example of asymptotic escape on the low-rank matrix manifold

In this section, we consider the phase retrieval problem [24, 98, 122] on the rank-1 matrix manifold. This serves both as an application of our asymptotic escape analysis for strict saddles, and as a demonstration of the possibility of treating such a problem rigorously on the manifold as opposed to the Euclidean space.

Since the phase retrieval problem involves a large number of stochastic measurements (i.e., random coefficient matrices $\{A_j(\omega)\}$ that constitute the objective function f_ω, ω indicating the random event), we will approach this problem in two steps. First, a crude analysis will be performed on its expectation $\mathbb{E}_\omega f$. In this case we will locate a strict critical submanifold in the shape of a "hyper ring". Then, for the non-expectation case $f = f_{(\omega)}$, we will prove a rather surprising result that it almost surely has only a finite number of saddle points. We will then show that with high probability, these saddles are strict saddles, and we know they are located near the above "hyper ring", so our asymptotic escape analysis is also applicable. The asymptotic escape is further supported by numerical experiments.

Phase retrieval on manifold: the expectation

The problem of phase retrieval in the case of real values aims to retrieve the information about $x \in \mathbb{R}^n$, from the phaseless measurements

$$y_j = |a_j^\top x|^2, \quad j = 1, \dots m,$$

where the entries of $\{a_j(\omega)\}_{j=1}^m$ are drawn from i.i.d. Gaussian. Usually a large m is needed to ensure successful recovery of x.

Let $X = xx^\top$, $A_j = a_j a_j^\top$, then $y_j = \langle A_j, X\rangle$. The problem can be posed on the rank-1 matrix manifold $\mathcal{M}_1$ as

$$\min_{Z \in \mathcal{M}_1} f(Z) := \frac{1}{2m} \sum_{j=1}^{m} |\langle A_j, Z - X\rangle|^2.$$

We can apply the Riemannian gradient descent to solve this problem on $\mathcal{M}_1$. We refer the reader to [24] in which the authors discussed the practical aspects of the RGD algorithm applied to phase retrieval. It is easily seen that $Z = X$ is the unique global minimizer. To ensure asymptotic convergence of the RGD to the global minimizer, it remains to rule out local minimizers and identify other critical points as strict saddles. Previous works [98, 122] have shown that phase retrieval has no spurious local minimum at least with high probability in the Euclidean setting. The analysis of saddle has been more complicated because of the stochasticity and Euclidean space parameterization.

It helps to take the expectation of $f(Z)$ and look into its landscape on the manifold. Note that

$$\bar{f}(Z) := \mathbb{E}_\omega f(Z) = \frac{3}{2}\|Z\|_F^2 + \frac{3}{2}\|X\|_F^2 - \|Z\|_F\|X\|_F - 2\langle Z, X\rangle,$$

and the Riemannian gradient (i.e., projected gradient) is

$$\mathrm{grad}\bar{f}(Z) = P_{T_Z}(\nabla\bar{f}(Z)) = P_{T_Z}((3 - \frac{\|X\|_F}{\|Z\|_F})Z - 2X).$$

The first-order condition is satisfied if either $Z = X$, or

$$\|Z\|_F = \frac{1}{3}\|X\|_F, \quad \langle Z, X\rangle = 0.$$

The latter are spurious critical points, and they form a (n-2)-dimensional submanifold on $\mathcal{M}_1$. To see whether they are strict saddles, we look into their Hessian.

Let $Z = zz^\top$, $u = z/\|z\|_2$, then $u \perp x$. Any element $\xi \in T_Z$ can be represented as $\xi = wuu^\top + uv^\top + vu^\top$, where $w \in \mathbb{R}$, $v \in \mathbb{R}^n$ and $v \perp u$. From [125], $R_N(Z+\xi) = Z + \xi + \eta + O(\|\xi\|^3)$ where $\eta = vv^\top/\|Z\|_F$. Using the formula that $\mathrm{Hess}f(Z) = \mathrm{Hess}(f \circ \mathcal{R}_Z)(t\xi)\mid_{t=0}$, we have

$$\begin{aligned} f \circ \mathcal{R}_Z(\xi) &= f(Z + \xi + \eta) + O(\|\xi\|^3) \\ &= f(Z) + \langle\nabla f(Z), \xi\rangle + \langle\nabla f(Z), \eta\rangle + \frac{1}{2}\langle\nabla^2 f(Z)[\xi], \xi\rangle + O(\|\xi\|^3), \end{aligned}$$

and collecting the second-order terms gives

$$\begin{aligned} &\langle\mathrm{Hess}\bar{f}(Z)[\xi], \xi\rangle = 2\langle\nabla\bar{f}(Z), \eta\rangle + \langle\nabla^2\bar{f}(Z)[\xi], \xi\rangle \\ &= (6 - 2\frac{\|X\|_F}{\|Z\|_F})\langle Z, \eta\rangle - 4\langle X, \eta\rangle + (3 - \frac{\|X\|_F}{\|Z\|_F})\|\xi\|_F^2 + \frac{\|X\|_F}{\|Z\|_F^3}\langle Z, \xi\rangle^2 \\ &= -4\langle X, \eta\rangle + \frac{3}{\|Z\|_F^2}\langle Z, \xi\rangle^2 \\ &= -4\frac{|x^\top v|^2}{\|Z\|_F} + 3w^2. \end{aligned}$$

Let $\xi = ux^\top + xu^\top$, then $\langle \mathrm{Hess}\bar{f}(Z)[\xi], \xi\rangle = -12\|X\|_F < 0$. Therefore these spurious critical points are strict saddles. In fact they form a strict critical submanifold $\mathcal{N} = \{Z \in \mathcal{M} \mid \|Z\|_F = \frac{1}{3}\|X\|_F, \quad \langle Z, X\rangle = 0\}$. For $p \in \mathcal{N}$, $T_p\mathcal{M} = dim(T_p\mathcal{N}) = n-2$, $dim(\nu_p^s\mathcal{M}) = dim(\nu_p^u\mathcal{M}) = 1$. RGD will escape the strict critical submanifold and converge to the minimum of $\bar{f}$ almost surely by Theorem 3.3.4.

Note that although we focus on the real case (i.e., $\mathcal{M}_1(\mathbb{R})$) here, the above results can be generalized to the complex case easily, and the only change is in the constants concerning Gaussian moments.

Phase retrieval: Dive into specific realizations

Specific realizations of phase retrieval may have much more complicated landscape than the expectation case. However, in the previous work [98] the authors have shown that for a slightly modified objective function, with high probability, the saddles of a specific realization of phase retrieval lie in the neighborhood of the above $\mathcal{N}$, what we call the "hyper ring".

Consider

$$f(Z) = \frac{1}{2m}\sum_{j=1}^{m} |\langle A_j, Z - X\rangle|^2$$

for a specific realization of $\{A_j(\omega)\}_{j=1}^m$. The Riemannian gradient is

$$\mathrm{grad} f(Z) = P_{T_Z}(\nabla f(Z)) = \frac{1}{m}\sum_{j=1}^{m}\langle A_j, Z - X\rangle P_{T_Z}(A_j).$$

And the Riemannian Hessian is

$$\begin{aligned}\langle \mathrm{Hess} f(Z)[\xi], \xi\rangle &= 2\langle \nabla f(Z), \eta\rangle + \langle \nabla^2 f(Z)[\xi], \xi\rangle \\ &= \frac{1}{m}\sum_{j=1}^{m}(2\langle A_j, Z - X\rangle\langle A_j, \eta\rangle + \langle A_j, \xi\rangle^2).\end{aligned}$$

The first result is a rather surprising one showing the finite number of critical points for phase retrieval.

Lemma 3.4.1. *When $m \geq n$, the above $f(Z)$ almost surely has only finite number of critical points on the manifold $\mathcal{M}_1$.*

The proof of Lemma 3.4.1 is quite neat using the following result from [62] and restated in [93].

Lemma 3.4.2. *For a polynomial system $P(x) = (p_1(x), \dots, p_n(x))$ with $x = (x_1, \dots x_n)$ and d_i = degree $p_i(x)$, let $p_i(x) = p_i^1(x) + p_i^2(x)$ where $p_i^1(x)$ consists of all the terms of $p_i(x)$ with degree d_i. If the homogeneous polynomial system $P^1(x) = (p_1^1(x), \dots, p_n^1(x)) = 0$ has only the trivial solution $x = 0$, then the original system $P(x) = 0$ only has a finite number of solutions. Moreover, the number of solutions is exactly $\Pi_{i=1}^n d_i$.*

Proof of Lemma 3.4.1. The first-order condition $\text{grad} f(Z) = 0$ is equivalent to

$$\frac{1}{m}\sum_{j=1}^{m} \langle A_j, Z - X\rangle P_{T_Z}(A_j) = 0.$$

Let $\tilde{U} \in \mathbb{R}^{n\times(n-1)}$ be the orthonormal complement of u. Then we have that

$$\begin{aligned} P_{T_Z}(A_j) &= uu^\top A_j uu^\top + uu^\top A_j \tilde{U}\tilde{U}^\top + \tilde{U}\tilde{U}^\top A_j uu^\top \\ &= u(a_j^\top u)^2 u^\top + u(a_j^\top u \cdot a_j^\top \tilde{U})\tilde{U}^\top + \tilde{U}(a_j^\top u \cdot \tilde{U}^\top a_j)u^\top. \end{aligned}$$

Applying a basis transform $(u, \tilde{U})$ to the first-order condition, by symmetry, it is equivalent to

$$\begin{cases} \frac{1}{m}\sum_{j=1}^m \langle A_j, Z - X\rangle \cdot a_j^\top u \cdot a_j^\top u = 0, \\ \frac{1}{m}\sum_{j=1}^m \langle A_j, Z - X\rangle \cdot a_j^\top u \cdot a_j^\top \tilde{U} = 0, \end{cases}$$

which is equivalent to $\sum_{j=1}^m \langle A_j, Z - X\rangle (a_j^\top u) a_j = 0$, i.e., finding $z \in \mathbb{R}^n$ such that

$$\sum_{j=1}^{m} (|a_j^\top z|^2 - |a_j^\top x|^2)(a_j^\top z) a_j = 0. \tag{3.2}$$

This is a third-order heterogeneous polynomial system of n equations for n unknowns. The homogeneous part of the system is

$$\sum_{j=1}^{m} |a_j^\top z|^2 (a_j^\top z) a_j = 0.$$

This system almost surely only has the trivial solution $z = 0$. To see this, note that it requires $\sum_{j=1}^m |a_j^\top z|^4 = 0$, i.e.

$$a_j^\top z = 0, \quad j = 1, \dots, m.$$

Since $\{a_j\}$ are i.i.d. Gaussian, when $m \geq n$ this linear system is almost surely nondegenerate. Now we can apply Lemma 3.4.2 and deduce that the original system only has finite number of solutions, i.e., $f(Z)$ only has finite number of critical points on the manifold. □

Remark 3.4.3. From Equation (3.2), we can see that the first-order condition on the manifold $\mathcal{M}_1$ is equivalent to that in the parameterized Euclidean space. This means that their critical points match. Still, a critical point $Z = zz^\top$ corresponds to at least two critical points $\pm z$ in the parameterized Euclidean space. Also, their Hessian can be very different.

Remark 3.4.4. The result of Lemma 3.4.1 only applies to the case $z \in \mathbb{R}^n$. In the case $z \in \mathbb{C}^n$, we conjecture that there would be a finite number of of critical submanifolds instead. Each critical submanifold consists of $\{e^{i\theta}z_* : \theta \in [0, 2\pi)\}$, the family of *phaseless* vectors. To see this, we can impose the constraints $a_j^H z \in \mathbb{R}$ to the above equations (this is always possible by letting a_j absorb the phase information, which does not alter A_j). Now we can replace $|\cdot|$ with $(\cdot)$ and again get a polynomial system. Lemma 3.4.2 is still applicable, and we get the finiteness of solutions on this constrained subset. To remove the constraints, we put the phase information back and obtain the submanifolds.

The Hessians of saddle points in phase retrieval are treated in the next lemma. Note that the condition $m \geq n$ in Lemma 3.4.1 only ensures the finite number of saddles. To make sure that saddles are strict, we need $m \gtrsim n \log n$, which is consistent with recovery guarantees from previous works (see e.g., [24] and references therein).

Theorem 3.4.5. *Given $\delta_0, \delta_1 > 0$. If $m \geq C(\delta_1) n \log n$, then with high probability no less than $1 - \frac{C_1}{m} - e^{-C_2 n}$, for all Z that satisfy the following conditions*

$$\begin{cases} \langle Z, X \rangle \leq \delta_0 \|Z\|_F \|X\|_F, \\ \frac{1}{3} - \delta_0 \leq \frac{\|Z\|_F}{\|X\|_F} \leq \frac{1}{3} + \delta_0, \\ P_{T_Z}(\nabla f(Z)) = 0, \end{cases}$$

we have

$$\lambda_{min}(Hess\ f(Z)) \leq \Lambda(\delta_0, \delta_1) < 0.$$

Here C_1, C_2 are absolute constants, $C(\delta_1)$ depend only on δ_1, and Λ depend only on δ_0 and δ_1. If we further require $\delta_0 < \frac{1}{6}$, $\delta_1 < \frac{5}{36}$, then $\lambda_{min}(Hess f(Z)) < -1$.

Proof of Theorem 3.4.5. The construction of a negative curvature direction is similar to that in the previous subsection. Let $\xi = xu^\top + ux^\top$, then $\xi \in T_Z$.

Since now x and z are not orthogonal, $\xi = wuu^\top + uv^\top + vu^\top$, where $w = 2u^\top x$ and $v = x - uu^\top x$. The Hessian is

$$\begin{aligned}\langle \mathrm{Hess} f(Z)[\xi], \xi\rangle &= \frac{1}{m}\sum_{j=1}^{m}(2\langle A_j, Z - X\rangle\langle A_j, \eta\rangle + \langle A_j, \xi\rangle^2) \\ &= \frac{1}{m}\sum_{j=1}^{m}(2\langle A_j, Z - X\rangle\langle A_j, \frac{xx^\top}{\|Z\|_F} + (\eta - \frac{xx^\top}{\|Z\|_F})\rangle + \langle A_j, \xi\rangle^2).\end{aligned}$$

An important observation is

$$\frac{1}{m}\sum_{j=1}^{m}(2\langle A_j, Z - X\rangle\langle A_j, (\eta - \frac{xx^\top}{\|Z\|_F})\rangle = 0.$$

This is because

$$\begin{aligned}\eta \cdot \|Z\|_F - xx^\top = vv^\top - xx^\top &= (x - uu^\top x)(x - uu^\top x)^\top - xx^\top \\ &= -uu^\top xx^\top - xx^\top uu^\top + uu^\top xx^\top uu^\top \in T_Z,\end{aligned}$$

and the first-order condition gives $\frac{1}{m}\sum_{j=1}^{m}\langle A_j, Z - X\rangle\langle A_j, \zeta\rangle = 0$ for any $\zeta \in T_Z$.

Therefore, we have

$$\begin{aligned}\frac{\langle \mathrm{Hess} f(Z)[\xi], \xi\rangle}{\|\xi\|_F^2} &= \frac{\frac{1}{m}\sum_{j=1}^{m}(2\langle A_j, Z - X\rangle\langle A_j, \frac{xx^\top}{\|Z\|_F}\rangle + \langle A_j, \xi\rangle^2)}{\|\xi\|_F^2} \\ &= \frac{\frac{1}{m}\sum_{j=1}^{m}(2(|a_j^\top z|^2 - |a_j^\top x|^2)|a_j^\top x|^2 + 4|a_j^\top z|^2|a_j^\top x|^2)}{2(\|z\|^2\|x\|^2 + \langle x, z\rangle^2)} \\ &= \frac{\frac{1}{m}\sum_{j=1}^{m}(3|a_j^\top z|^2|a_j^\top x|^2 - |a_j^\top x|^4)}{\|z\|^2\|x\|^2 + \langle x, z\rangle^2}.\end{aligned}$$

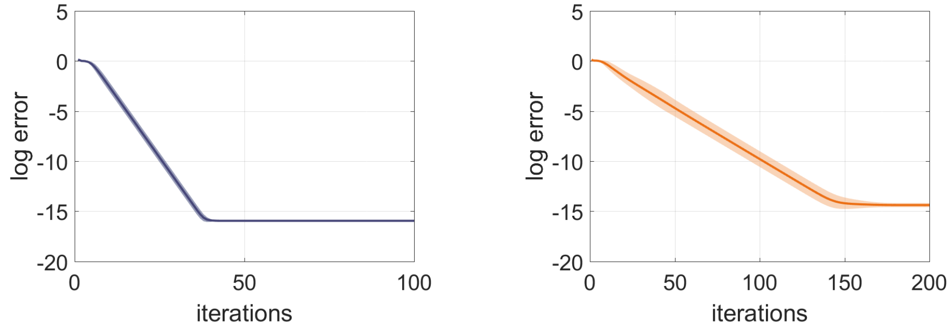
Using the concentration inequalities from Section 4 in [77], with high probability no less than $1 - \frac{C_1}{m} - e^{-C_2 n}$, we have

$$\begin{aligned}\frac{\langle \mathrm{Hess} f(Z)[\xi], \xi\rangle}{\|\xi\|_F^2} &\le \frac{3(1+\delta_1)(\|Z\|_F\|X\|_F + 2\langle X, Z\rangle) - (3 - \delta_1)\|X\|_F^2}{\|Z\|_F\|X\|_F + \langle X, Z\rangle} \\ &\le \frac{3(1+\delta_1)(\frac{1}{3} + \delta_0 + 2\delta_0(\frac{1}{3} + \delta_0)) - (3 - \delta_1)}{(\frac{1}{3} + \delta_0) + \delta_0(\frac{1}{3} + \delta_0)} := \Lambda(\delta_0, \delta_1).\end{aligned}$$

If $\delta_0 < \frac{1}{6}$, $\delta_1 < \frac{5}{36}$, then we get $\Lambda(\delta_0, \delta_1) < -1$. □

The above results give us a good idea of the critical points in the "hyper ring" region $\{\frac{1}{3} - \delta_0 \le \frac{\|Z\|_F}{\|X\|_F} \le \frac{1}{3} + \delta_0\}$ on the manifold. Specifically, Lemma

3.4.1 tells us that there are only a finite number of critical points, and Theorem 3.4.5 asserts that these critical points are all strict saddles on the manifold since they have a common negative curvature direction. We are particularly interested in the "hyper ring" region because Theorem 2.2 of [98] shows (with a slightly modified objective function) that all the critical points lie in this region with high probability, except the unique global minimum. From Theorem 3.2.3, we now know that the RGD will avoid saddles and converge to the global minimum.



(a) $log10$ error for the expectation case. (b) $log10$ error for a specific realization.

Figure 3.2: Convergence (visualized as error band) of RGD for phase retrieval.

Figure 3.2 shows the $log10$ error convergence of the RGD for phase retrieval on the manifold $\mathcal{M}_1$. The left figure is about the expectation case, also called the population problem, while the right one is a specific realization with a certain group of $\{A_j\}_{j=1}^m$, where $m = 12n$. In both experiments, we take $n = 256$, learning rate $\alpha = \frac{1}{3}$, draw 100 z_0 from i.i.d. Gaussian distribution ($Z_0 = z_0 z_0^\top$), and minimize $\mathbb{E}f(Z)$ or $f(Z)$ starting from these random initializations. The darker central line is the average, and the band shows the deviation. In general, it can be seen that the RGD is hardly affected by the possible existence of saddle points and converges to the minimum.

This experiment has also demonstrated the curious phenomenon mentioned at the beginning of Section 2.1, namely a first-order method such as RGD converges exponentially fast (i.e., linearly), even though in the Euclidean space it does not (i.e., only sublinearly). This will be explained in upcoming Chapter 5.

3.5 Example of applications on other Riemannian manifolds

Although our primary setting is the low-rank matrix manifold $\mathcal{M}_r$, the asymptotic convergence to the minimum and escape of strict saddles (strict critical submanifolds) is valid on arbitrary finite dimensional Riemannian manifold $\mathcal{M}$. In particular, the properties of the RGD are well preserved if the manifold is embedded in a Banach space and inherits the Riemannian metric from this ambient space. Below we discuss the optimization on the unit sphere and the Stiefel manifold as two examples of applications.

Variational eigenproblem on a sphere

Consider $\mathcal{M} = \mathbb{S}^{n-1}$, the sphere embedded in the Euclidean space $\mathbb{R}^n$. We consider the following eigenvalue problem:

$$g(z) = \lambda z, \quad z \in \mathbb{R}^n,$$

where $g(z)$ is a function of z that may or may not be linear in z. Assume that it has eigenpairs $(\lambda_1, v_1), (\lambda_2, v_2), \ldots, (\lambda_k, v_k)$, $0 < \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_k$. If $g(z) = \nabla f(z)$ for some function $f(z)$, then to find (λ_1, v_1) is to solve the following optimization problem:

$$\min_z f(z) \quad \text{s.t. } z \in \mathcal{M} = \mathbb{S}^{n-1}.$$

Viewed as an embedded Riemannian manifold, the tangent space, tangent space projection, and retraction on $\mathcal{M} = \mathbb{S}^{n-1}$ are given as follows:

$$\begin{aligned} T_z &= \{\xi \in \mathbb{R}^n : \xi^\top z = 0\}, \\ P_{T_z} &= I - zz^\top, \\ R(y) &= \frac{y}{\|y\|_2}. \end{aligned}$$

Note that $R(y)$ is a second-order retraction, because for any $z \in \mathcal{M}$, $\xi \in T_z$, we have

$$R(z + \alpha\xi) = \frac{z + \alpha\xi}{\|z + \alpha\xi\|_2} = (z + \alpha\xi)(1 + \alpha^2\|\xi\|_2^2)^{-\frac{1}{2}} = z + \alpha\xi + O(\alpha^2).$$

The Levi-Civita connection on $\mathcal{M}$ is the projection of the Levi-Civita connection of the ambient space (which is the derivative in $\mathbb{R}^n$)

$$\widetilde{\nabla}_{\xi_z}\eta = P_{T_z}(\nabla_{\xi_z}\eta) = (I - zz^\top)(\nabla_{\xi_z}\eta), \quad \eta \in T_{\mathcal{M}}, \quad \xi_z \in T_z.$$

The Riemannian gradient on $\mathcal{M}$ is

$$\mathrm{grad} f(z) = P_{T_z}(\nabla f(z)).$$

So z is a critical point on $\mathcal{M}$ if and only if z is an eigenvector of the eigenproblem $g(z) = \lambda z$. The Riemannian Hessian on $\mathcal{M}$ is

$$\mathrm{Hess} f(z)[\xi] = P_{T_z}(\nabla^2 f(z)[\xi]) - (z^\top \nabla f(z))\xi.$$

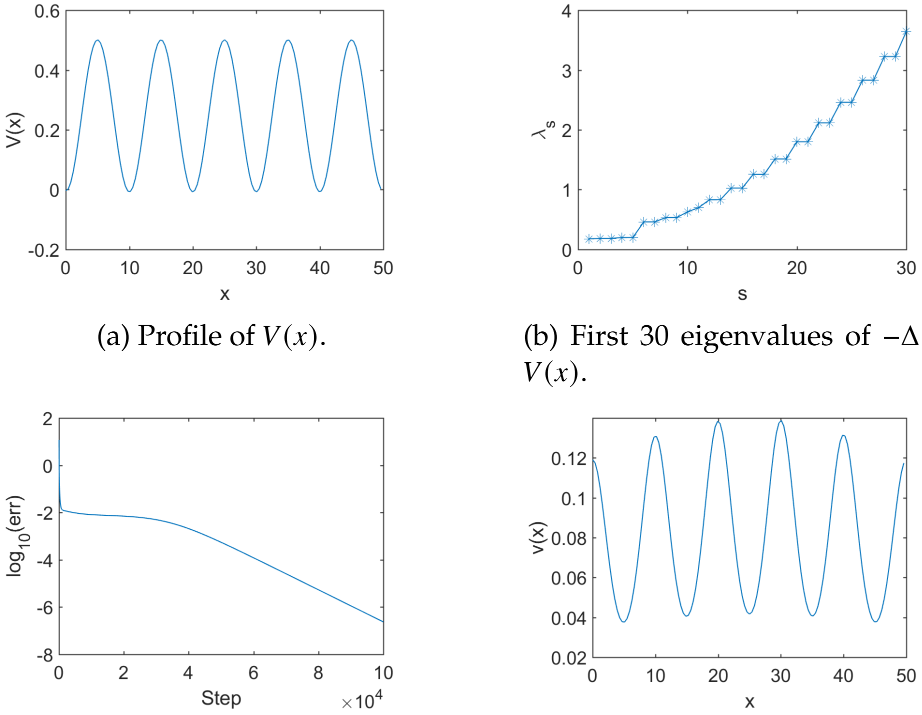
If $g(z)$ is linear in z, then $f(z)$ is quadratic. With the positiveness assumption, we have $f(z) = z^\top A z$, where A is an SPD matrix. Then $f(x_i) = \lambda_i$, $\mathrm{grad} f(z) = Az - (z^\top A z)z$, and $\xi^\top \mathrm{Hess} f(z)[\xi] = \xi^\top A \xi - (z^\top A z)\xi^\top \xi$. It is easy to see that v_1 is the unique (up to sign) global minimum with a positive Hessian, and $v_s (s > 1)$ are all strict saddles whose Hessian has at least one negative curvature direction $\xi = v_s$.

It is interesting to look at the case where a non-minimal eigenvalue has multiplicity greater than 1. Assume that $\lambda_s = \lambda_{s+1} = \ldots = \lambda_{s+t}$, then the submanifold $\mathcal{N} = \{y \in \mathbb{R}^n \mid y = c_s v_s + \ldots + c_{s+t} v_{s+t}, \quad c_s^2 + \ldots + c_{s+t}^2 = 1\}$ is an immersed submanifold of $\mathcal{M}$, and it is a strict critical submanifold of f if $s \geq 2$. Since the number of such submanifolds is finite, escape from these submanifolds toward x_1 is ensured by the tools in Section 3.3.

When $g(z)$ is not linear in z, as $f(z)$ now contains non-quadratic terms, it is not immediately clear from the algebraic expression whether $\mathrm{Hess} f(x_s), s > 1$ has negative curvature direction, though it can be verified numerically.

The first numerical example is from the discretized 1D Schrödinger eigenproblem $-\Delta u + V(x)u = \lambda u$ with periodic boundary condition, where $V(x)$ is taken to be the smoothed 1D Kronig-Penney (KP) potential describing free electrons in 1D crystal [87, 111]. Figure 3.3a shows the profile of the KP potential defined on $D = [0, 50]$ with 5 energy wells and periodic BC. Figure 3.3b shows the first 30 eigenvalues of the operator $-\Delta + V(x)$. We can see that the first 5 eigenvalues are clustered (but not identical).

We discretize D into $n = 2^7$ grids and solve the discretized problem on $\mathcal{M} = \mathbb{S}^{n-1}$ with the RGD starting from a random initialization. The step size is $\alpha = 0.01$. In Figure 3.3c, we observe that the generated sequence first seems to "stagnate" near a non-minimal eigenstate, but then escapes and converges toward the minimum. Figure 3.3d shows the profile of the computed ground energy state v_1, which is quite close to the true ground state but slightly deformed. An improvement will be proposed in the next subsection.



(a) Profile of $V(x)$.

(b) First 30 eigenvalues of $-\Delta + V(x)$.

(c) Error decay in RGD.

(d) Profile of the first eigenstate v_1.

Figure 3.3: Solving the linear Schrödinger eigenproblem on the sphere.

The second example is the nonlinear Schrödinger eigenproblem $-\Delta u + V(x)u + \beta|u|^2 u = \lambda u$, or the so-called Gross-Pitaevskii eigenvalue problem for the Bose-Einstein Condensate (BEC) [115]. It gives a more accurate description of the dynamics of Bosonic gases at ultra-low temperature. With the presence of the nonlinear term $\beta|u|^2 u$, linear eigensolvers would fail, and the optimization of its variational form becomes the state-of-art solver, see e.g., [72]. Apart from the RGD based on the L^2 metric, there can be other RGD algorithms based on other types of metrics and with different convergence theories, whose analysis is beyond the scope of this chapter.

We use the same potential function $V(x)$ and discretization size as above.

The nonlinear term has the weight $\beta = 1$. The objective function is

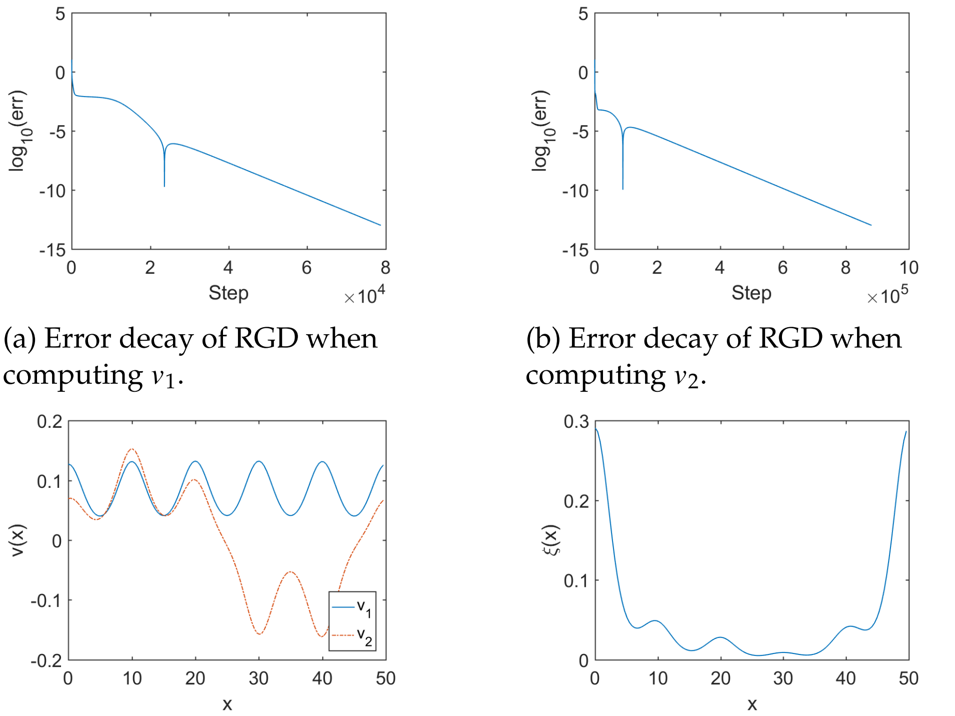
$$f(z) = \frac{1}{2} z^\top A z + \frac{\beta}{4} \sum_{j=1}^{n} z(j)^4, \quad A = -L + V.$$

For an eigenstate v_s, the eigenvalue associated to it is

$$\lambda_s = 2f(s) + \frac{\beta}{2} \sum_{j=1}^{n} z(j)^4.$$

We compute the first two eigenstates of the nonlinear Schrödinger problem using the RGD with stepsize $\alpha = 0.01$. Figure 3.4c shows their profiles. Figures 3.4a and 3.4b demonstrate the convergence of the RGD toward the computed eigenvalues.

To verify that v_2 is a strict saddle point, we numerically compute the smallest eigenvalue of $\mathrm{Hess} f(v_2)$ and $\lambda_{\min}(\mathrm{Hess} f(v_2)) = -0.0024 < 0$. Figure 3.4d shows a profile of the corresponding eigenvector $\xi_{\min}$, i.e., a negative curvature direction.



(a) Error decay of RGD when computing v_1.

(b) Error decay of RGD when computing v_2.

(c) Profile of eigenstates v_1 and v_2.

(d) A negative curvature direction.

Figure 3.4: Solving the nonlinear Schrödinger eigenproblem on the sphere.

Apart from physical problems like BEC eigenstates, linear and nonlinear eigenproblems also find applications in image processing and machine learning. For example, the MaxCut problem corresponds to a linear eigenproblem, while the optimization of the Ginzburg-Landau type functional in image segmentation and learning tasks corresponds to a nonlinear eigenproblem, see e.g., [19, 75]. Although there are many algorithms tailored for linear eigenproblems, their nonlinear relatives often lack a rigorous convergence guarantee. Manifold optimization thus provides a more versatile point of view for them.

Simultaneous eigensolver on the Stiefel manifold

Subspace iteration is a common technique for accelerating the convergence of smallest eigenstates in linear eigenproblems, especially when the ground states are clustered, as is in the previous examples.

From the viewpoint of manifold optimization, to solve the first m eigenstates simultaneously can be posed as the optimization on the Stiefel manifold $\mathcal{M} = \{Z \in \mathbb{R}^{n\times m} : Z^\top Z = I_m\}$:

$$\min_Z \ \mathrm{trace}(f(Z)) \quad \text{s.t. } Z \in \mathcal{M} = \{Z \in \mathbb{R}^{n\times m} : Z^\top Z = I_m\}.$$

The Stiefel manifold [53, 83] is the set of all m-frames in $\mathbb{R}^n$. When $m = 1$, it reduces to the sphere $\mathbb{S}^{n-1}$. With the Euclidean metric, its tangent space, tangent space projection and retraction are given as follows:

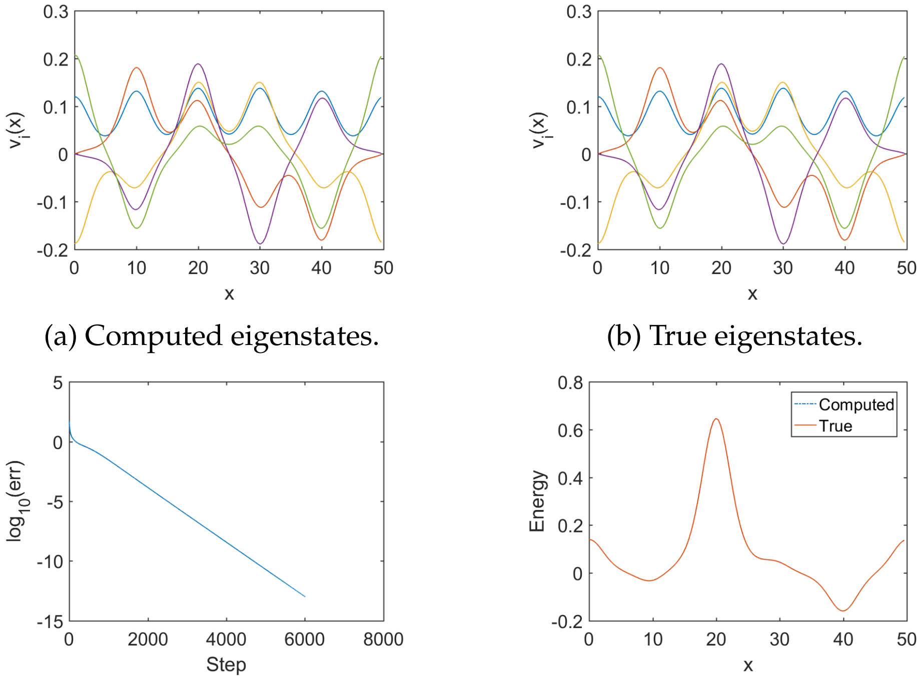
$$\begin{aligned} T_Z &= \{\xi \in \mathbb{R}^{n\times m} : \xi^\top Z + Z^\top \xi = 0\}, \\ P_{T_Z}(Y) &= Y - Z\,\mathrm{sym}(Z^\top Y), \\ R(Y) &= \mathrm{qf}\,(Y), \end{aligned}$$

where $\mathrm{sym}(\cdot)$ takes the symmetric part and $\mathrm{qf}(\cdot)$ takes the Q factor of QR decomposition. Similar to the case of the sphere manifold, the Riemannian connection and gradient are defined by the projection onto the tangent space. When $f(Z) = Z^\top AZ$, we have

$$\begin{aligned} \mathrm{grad} f(Z) &= P_{T_Z}(AZ), \\ \langle \xi, \mathrm{Hess} f(Z)[\xi] \rangle &= \mathrm{tr}(\xi^\top A\xi - (\xi^\top \xi)(Z^\top AZ)). \end{aligned}$$

It is easily verified that the minimum is achieved when $\mathrm{span} Z = \mathrm{span}\{v_1, \ldots, v_m\}$, and all Z that span other eigen subspaces are strict saddles if all the eigenvalues are distinct.

We compute the first 5 eigenstates simultaneously for the linear Schrödinger eigenproblem with the same potential as in Figure 3.3a. The step size is $\alpha = 0.01$. Figures 3.5a and 3.5b compare the computed eigenstates extracted from Z and the true eigenstates, which are almost identical. In Figure 3.5c, we can see that the subspace iteration on the Stiefel manifold achieves much better convergence in fewer steps than the optimization on the sphere.



(a) Computed eigenstates.

(b) True eigenstates.

(c) Error decay of RGD.

(d) Accumulated energy of the first 5 eigenstates.

Figure 3.5: Simultaneously solving the first 5 eigenstates of the linear Schrödinger problem on the Stiefel manifold.

Application of the Stiefel manifold optimization can also be extended to data science, e.g., frame construction and dictionary learning [25, 121], if the frame/dictionary satisfies orthonormal assumptions.

3.6 Discussion

We have studied the asymptotic escape of strict saddles points of the RGD on the Riemannian manifolds. The first main contribution of this chapter is that it pushes the boundary of current analysis to non-isolated saddle sets, proving when the RGD can escape and indicating when it cannot. As a general tool, it can be applied to various settings as long as the manifold of interest satisfies certain smoothness conditions. This is demonstrated by several representative examples from different fields.

The saddle analysis of phase retrieval on the low-rank matrix manifold serves as an application of the above asymptotic escape result, but it also stands as an insightful result by itself. We have shown that it always has a finite number of critical points, and the saddles are strict saddles with high probability. Essentially, the low-rank matrix manifold sheds light on the intrinsic quadratic (instead of quartic) structure of this problem.

In addition to the asymptotic convergence behavior of the RGD, the convergence rate is also an important issue. Empirical linear convergence rates in many low-rank matrix recovery problems are already observed but are yet to be explained. This is the topic of Chapter 5, where we prove the linear convergence rate using the quadratic nature of those problems on the manifold $\overline{\mathcal{M}_r}$.

3.7 Appendix

As we mentioned in Section 3.3, when there are a bunch of self-connected critical submanifolds (generalization of critical points), the escape of strict critical submanifolds (generalized strict saddles) and convergence to a minimum rely on the number or the structure of such critical submanifolds. When the number is uncountable, the situation can be quite complicated.

In this appendix, we discuss some structural properties of critical submanifolds that may help untangle their successive relations. We introduce the concepts of index and transversality, point out the transversality properties of certain functions and their consequences, and link the stable manifolds of the gradient flow to that of the gradient descent.

Definition 3.7.1 (Index)**.** For $f : \mathcal{M} \mapsto \mathbb{R}$, let p be a critical point of f, then the *index* of p is

$$\lambda_p := \dim T_p^u \mathcal{M}.$$

Remark 3.7.2. All critical points in the same connected critical submanifold $\mathcal{N}$ have the same index, and we define it as the index $\lambda_{\mathcal{N}}$ of the submanifold $\mathcal{N}$. An equivalent way to define strict critical submanifold is $\lambda_{\mathcal{N}} > 0$.

Definition 3.7.3 (Transversality)**.** (1) For smooth maps $f : \mathcal{N}_1 \mapsto \mathcal{M}$ and $g : \mathcal{N}_2 \mapsto \mathcal{M}$, we say that f is *transverse* to g, iff for any X_1, X_2 such that

$f(X_1) = g(X_2) = Y$,

$$df(T_{X_1}\mathcal{N}_1) + dg(T_{X_2}\mathcal{N}_2) = T_Y\mathcal{M},$$

where df and dg are gradient vector fields of f and g;
(2) If $\mathcal{N}_1$ and $\mathcal{N}_2$ are immersed submanifolds of $\mathcal{M}$, then $\mathcal{N}_1$ is *transverse* to $\mathcal{N}_2$ iff for any $X \in \mathcal{N}_1 \cap \mathcal{N}_2$,

$$T_X\mathcal{N}_1 + T_X\mathcal{N}_2 = T_X\mathcal{M}.$$

Two immersed submanifolds vacuously transverse if they do not intersect.

Remark 3.7.4. A function $f : \mathcal{M} \mapsto \mathbb{R}$ is called *Morse-Bott* if all its critical points lie in some disjoint union of connected and nondegenerate critical submanifolds; f is called *Morse-Smale* if it satisifies the Morse-Smale transversality condition, i.e., for any two critical submanifolds $\mathcal{N}_1$, $\mathcal{N}_2$, their stable and unstable manifolds intersect transversally.

The transversality condition for immersed manifolds simply means that two manifolds "cross" each other and do not "overlap". Figure 3.6 is a vivid illustration of transversality on a 2-dimensional manifold. If the objective function f is a Morse-Smale function, transversality implies more favorable properties.

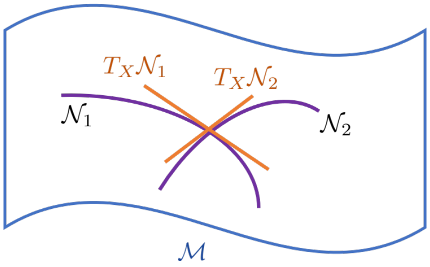


Figure 3.6: An illustration of transversality.

Theorem 3.7.5 (Corollary 6.27 in [13]). *For a Morse-Smale function f, any critical point p of f satisfies*

$$\overline{W^u(p)} = \bigcup_{p \geq q} W^u(q),$$
$$\overline{W^s(p)} = \bigcup_{r \geq p} W^s(r),$$

where $W^s(p)$ (resp. $W^u(p)$) is the stable (resp. unstable) manifold of p defined by gradient flow, and $p \geq q$ means $W^u(p) \cap W^s(q) \neq \emptyset$.

Theorem 3.7.6. *For a Morse-Smale function f, if two critical submanifolds $\mathcal{N}_1$ and $\mathcal{N}_2$ have the same index, then they vacuously transverse, i.e., $W^u(\mathcal{N}_1) \cap W^s(\mathcal{N}_2) = \emptyset$.*

Proof. By Proposition 6.2 in [13], if $W^u(\mathcal{N}_1) \cap W^s(\mathcal{N}_2) \neq \emptyset$, then their intersection is an embedded submanifold of dimension $(\lambda_{\mathcal{N}_1} - \lambda_{\mathcal{N}_2})$. But $\lambda_{\mathcal{N}_1} - \lambda_{\mathcal{N}_2} = 0$, which is a contradiction. □

Both Theorem 3.7.5 and Theorem 3.7.6 are helpful when taking the union of stable manifolds of infinitely many critical submanifolds. Theorem 3.7.5 shows that the closure of the stable/unstable manifold of one critical set is the union of the stable/unstable manifolds of the sets that have successive relations with it. On the other hand, Theorem 3.7.6 shows that the successive relations are strictly limited by the indices (i.e., negative curvature dimensions) of the critical sets. This successive relation simply cannot happen between sets of the same index.

It should be stressed that the above results are on the stable/unstable manifold of *gradient flows*, not *gradient descents*. Whether this can be generalized to gradient descents is still unclear. We know that with first-order retraction property, as $\alpha \to 0$, the Riemannian gradient descent on the manifold approximates the gradient flow trajectory. It can be proved that the respective stable/unstable manifolds also converge, as long as the retraction is at least first-order and the domain is compact. However, the transversality concerns the "angles" at the intersection of these submanifolds. Even the uniform convergence of submanifolds cannot ensure the preservation of their intersection angles along the convergence.

What lies at the core of the asymptotic analysis is an interesting interplay of dynamical systems and nonconvex optimization, and a translation of languages from the Morse theory [12, 13, 40] into gradient flows and further into gradient descents. Although these tools were initially developed to study homology, they have provided invaluable insight into the converging/escaping sets of strict saddle points with nontrivial geometry. This discussion aims to draw interest to the vast possibilities that Morse theory has to offer. They point out a way to deal with complex geometries of critical point sets. It would be interesting to further pursue this direction to quantify the above relations.

Chapter 4

ASYMPTOTIC ESCAPE OF SPURIOUS CRITICAL POINTS ON THE LOW-RANK MATRIX MANIFOLD

This chapter is focused on a fundamental problem that has long remained open in the analysis of the low-rank matrix manifold $\mathcal{M}_r$. It concerns the spurious critical points (Section 2.3), an intriguing phenomenon that is not present in the Euclidean space or other Riemannian manifolds. The spurious critical points $\mathcal{S}_\#$ are some rank-deficient matrices that capture part (but not all) of the eigen components of the ground truth. Unlike classical strict saddle points, they are singular and their Riemannian Hessian is unbounded. For a long time, people have used the low-rank matrix manifold $\mathcal{M}_r$ without realizing their existence. But we have seen in Section 2.3 that they could serve as limit points of certain minimizing sequences.

In this chapter, we show that the Riemannian gradient flow and the Riemannian gradient descent with a particular step size almost surely escape some spurious critical points on the boundary of $\mathcal{M}_r$. Our result is the first to partly overcome the non-closedness of the low-rank matrix manifold without changing the vanilla Riemannian gradient descent algorithm. Numerical experiments are provided to support our theoretical findings.

One important feature of the spurious critical points is that the Riemannian gradient around a spurious critical point is singular and the Riemannian Hessian is unbounded, see Figure 2.1 for an illustration. Because of this singularity, no asymptotic escape theorem in Chapter 3 can be applied directly to $\mathcal{S}_\#$. This poses a significant challenge to any attempt aiming at an asymptotic escape result.

To tackle this problem, the main technique we use is the dynamical low-rank approximation, which parameterizes the gradient flow on $\mathcal{M}_r$. By rescaling the gradient flow system in the parameterized domain, we are able to map a spurious critical point Z_* to a strict critical submanifold in the classical sense. This allows us to apply the result in Chapter 3 and show that Riemannian gradient descent and Riemannian gradient flow asymptotically escape rank-(r-1) spurious critical points.

The results in this chapter are purely asymptotic, in the sense that even though the algorithm escapes the spurious critical points, it is still unclear how long it takes to converge to the local minima. This question will be answered in Chapter 5, which establishes the linear convergence rate guarantee.

Organization of this chapter. We have given a brief introduction of the problem in Section 1.2. The rest of this chapter is organized as follows. In Section 4.1, we further elaborate on the background of the problem and related work. In Section 4.2 we present and prove the main result of this chapter, which is the asymptotic escape of the rank-(r-1) spurious critical points by the gradient flow. Specifically, we introduce the dynamical low-rank approximation as a primary tool, propose the rescaled gradient flow, prove its C^0- and C^1-extension to the rank-(r-1) spurious critical points, and show that these points are strict saddle points under the rescaled flow. In Section 4.3 we present the corresponding result for the gradient descent. In Section 4.4, some numerical experiments are performed to illustrate our theoretical results. Finally, Section 4.5 is devoted to some discussion.

Notations. Unless otherwise specified, upper-case letters stand for matrices, lower-case letters stand for vectors or scalars, and calligraphic letters stand for manifolds or sets. The field $\mathbb{F}$ can be either $\mathbb{R}$ or $\mathbb{C}$. The low-rank matrix manifold is denoted by $\mathcal{M}_r$, defined in Section 2.1. The Hermitian transpose is denoted by $(\cdot)^*$. The set of $n \times n$ real symmetric matrices is denoted by $\mathbb{S}_n$, while the set of $n \times n$ Hermitian matrices is denoted by $\mathbb{H}_n$. The Stiefel manifold is $\mathrm{St}(n, r) = \{U \in \mathbb{F}^{n\times r} : U^*U = I_r\}$. The orthogonal group is $\mathrm{SO}(n) = \mathrm{St}(n, n)$. The subscript $(\cdot)_\#$ is reserved for the spurious critical points. We use grad and Hess to denote the Riemannian gradient and Hessian, and ∇ to denote the Euclidean derivative.

4.1 Background and related work

We focus on the symmetric positive semi-definite (SPSD) or Hermitian positive semi-definite (HPSD) manifold. In Section 2.3, we have introduced the spurious critical points that emerge when minimizing the least squares loss function $f(Z) = \|Z - X\|_{\mathrm{F}}^2$ on $\mathcal{M}_r$. When it comes to the SPSD/HPSD case, one can define the spurious critical points as follows:

Definition 4.1.1 (Spurious critical points)**.** Assume that $X = U_X D_X U_X^*$ is an eigenvalue decomposition of $X \in \mathcal{M}_r$. Then the set of **spurious critical points** with respect to $f(Z) = \frac{1}{2}\|Z - X\|_{\mathrm{F}}^2$ on $\mathcal{M}_r$ is $\mathcal{S}_\# = \cup_{s=0}^{r-1}\mathcal{S}_s$, where each $\mathcal{S}_s$ can be characterized as

$$\mathcal{S}_s = \left\{Z_\# : \; Z_\# = U_1 D_1 U_1{}^*, \; U_1 \in \mathbb{F}^{n\times s}, \, D_1 \in \mathbb{F}^{s\times s}\right\}.$$

Here $U_X = (U_1, U_2)$, $U_1 \in \mathbb{F}^{n\times s}$, $U_2 \in \mathbb{F}^{n\times(r-s)}$ is a block decomposition of U_X; similarly for D_X.

The contribution of this chapter is to show that when minimizing the least squares loss function, the Riemannian gradient flow and the Riemannian gradient descent with varying step size asymptotically escape the rank-(r-1) spurious critical points on the rank-r SPSD or HPSD manifold, see Theorems 4.2.1 and 4.3.1.

In Section 1.2, we have introduced the motivation of our work. Below, we further review some related works in the literature and discuss their connection and comparison with our work.

Nonclosedness of the low-rank matrix manifold. The fact that $\mathcal{M}_r$ is not a closed set is first reported in [125] in the context of matrix completion, and later in [86] with low Tucker-rank tensor completion. To guarantee that the iterative sequence of the proposed algorithm stays inside a compact subset of $\mathcal{M}_r$, the author of [125] proposes to add a regularization term to the objective function f:

$$g(Z) = f(Z) + \mu^2(\|Z\|_{\mathrm{F}}^2 + \|Z^\dagger\|_{\mathrm{F}}^2),$$

where $Z^\dagger$ is the pseudo-inverse of Z, and μ is a parameter. In particular, the term $\mu^2\|Z^\dagger\|_{\mathrm{F}}^2$ guarantees that $\|Z^\dagger\|_{\mathrm{F}}$ will not go to infinity, i.e., the rank of Z will not drop below r.

However, the author also comments that μ^2 can be chosen very small, in fact as small as 10^{-16}. In numerical experiments, one can simply neglect this term and use the original function f instead of the regularized function g. In other words, the author has observed that even without regularization, the iterative sequence of the vanilla Riemannian gradient descent almost surely avoids the rank-deficient points and stays inside $\mathcal{M}_r$.

Apocalypses from a geometric point of view. Concurrent with our paper, the authors of [92] propose a similar concept. They use the term *apocalypse* to describe the event where the sequence of iterative points is in $\mathcal{M}_r$ but the limit point has rank less than r and is not stationary. This is exactly what happens in Example 2.3.2. They observe that apocalypse occurs when the tangent cone at the limit is not contained in the limit of the tangent cones. A more detailed discussion on the relation between tangent cones and optimality conditions can be found in [97].

Along this line of research, two remedies have been proposed to fix the apocalypse. The first is a second-order algorithm [92], which uses a smooth lift (similar to the Burer-Monteiro factorization) and the trust-region method. Another is a first-order algorithm proposed in [109], which uses the numerical rank to perform suitable rank reductions.

We remark that although both approaches avoid apocalyptic points, they require major modification to the gradient descent algorithm. In contrast, we focus on understanding why gradient descent needs no modification in practice, and we give a partial answer for the minimization of the least squares loss function.

Asymptotic escape of classical strict saddle points. We have discussed in Section 1.1 that gradient descent with random initialization almost surely escapes strict saddles and converges to minimizers. Results on the asymptotic escape of non-isolated strict saddle sets and strict critical submanifolds in the most general form are given in Chapter 3. The important observation is that the spurious critical points in $\overline{\mathcal{M}_r} \backslash \mathcal{M}_r$ are fundamentally different from, but subtly related to, the classical strict saddle points. The spurious critical points have singular local neighborhoods as illustrated in Figure 2.1. Their asymptotic escape behavior cannot be directly explained by Theorem 3.3.4. However, using a rescaled gradient flow, we can eliminate the singularity, and apply the saddle escape results to the rescaled system.

Implicit regularization in low-rank matrix factorization. The concept of *implicit regularization* is often used to describe the emergence of favorable structures without explicit regularization terms. In deep matrix factorization and deep neural networks, this describes a tendency toward low-rank solutions and better generalization [7]. In statistical estimation, this could mean a tendency to promote incoherence and accelerate convergence [39, 103]. As we have seen, the phenomenon that iterative sequences on the nonclosed manifold $\mathcal{M}_r$ stay inside the manifold does not rely on an explicit regularization term $\mu^2\|Z^\dagger\|_{\mathrm{F}}^2$. Thus it can also be seen as a form of implicit regularization.

Matrix decomposition and its continuity. Our analysis crucially relies on finding a low-rank decomposition that is sufficiently continuous along the whole gradient flow trajectory. The dynamical low-rank approximation (DLRA), first proposed in [85], is a decomposition that suits our purpose. In contrast, the singular value decomposition will lose its differentiability whenever singular values coalesce [49]. A variant called the *analytic SVD* [22] could fix this issue, but it requires analyticity of the gradient function. It cannot be applied to functions on $\mathcal{M}_r$ as the Riemannian gradient is not analytic because of the spurious critical points. We remark that the success of DLRA is still limited to the rank-(r-1) spurious critical points. Extension of the current analysis to general spurious critical points is left for future work.

4.2 Main result

Recall that by Definition 4.1.1, the set of spurious critical points of the least squares loss function on $\mathcal{M}_r$ is $\mathcal{S}_\# = \cup_{s=0}^{r-1}\mathcal{S}_s$, where each $\mathcal{S}_s$ $(0 \leq s \leq r-1)$ contains the rank-s spurious critical points, i.e.,

$$\mathcal{S}_s = \left\{Z_\# : \; Z_\# = U_1 D_1 U_1{}^*, \; U_1 \in \mathbb{F}^{n\times s}, \, D_1 \in \mathbb{F}^{s\times s}\right\}.$$

The first main result of this chapter is as follows.

Theorem 4.2.1 (Asymptotic escape of $\mathcal{S}_{r-1}$: gradient flow). *Let $f(Z) = \frac{1}{2}\|Z - X\|_{\mathrm{F}}^2$, where $X \in \mathcal{M}_r$ has distinct eigenvalues. Let $Z_t : t \geq 0$ be the gradient flow of f on $\mathcal{M}_r$ starting from a random initialization Z_0. Then we have that $Z_t \in \mathcal{M}_r$, $\forall\; 0 \leq t < +\infty$, and*

$$\mathit{Prob}\;(\lim_{t\to\infty} Z_t \in S_{r-1}) = 0.$$

In the next few sections, we introduce some technical tools that eventually constitute the proof of Theorem 4.2.1 at the end of Section 4.2.

Dynamical low-rank approximation

The dynamical low-rank approximation was first proposed in [85] and soon gained popularity as a discretization method for the computation of low-rank evolution systems. It gives a neat description of the column space and core matrix of the low-rank matrix along the evolution. The decomposition enjoys better smoothness than SVD and other classical decompositions. While a smooth version of SVD is only available when the gradient function is analytic, the dynamical low-rank approximation always preserves the smoothness of the gradient function. Below we first recall the general version of DLRA.

Lemma 4.2.2 (Dynamical low-rank approximation[1], [85])**.** *Consider the gradient flow of a function $f : \mathbb{F}_r^{m\times n} \to \mathbb{R}$, where $\mathbb{F}_r^{m\times n}$ is the set of $m\times n$ matrices with rank r. Assume $Z = USV^*$, where $U, V \in \mathbb{F}^{n\times r}$ are orthonormal, and $S \in \mathbb{R}^{r\times r}$ is nonsingular. Let $M := -\mathrm{grad} f(Z) = -P_{T_Z}(\nabla f(Z))$ denote the negative Riemannian gradient of f at $Z \in \mathcal{M}_r$. Impose the constraints $\dot{U}^*U = \dot{V}^*V = 0$. Then the gradient flow of f can be described by the following ODE system:*

$$\begin{cases} \dot{U} = P_U^{\perp} MVS^{-1}, \\ \dot{V} = P_V^{\perp} M^* U (S^{-1})^*, \\ \dot{S} = U^* MV. \end{cases} \tag{4.1}$$

Here, $P_U^{\perp} = I - UU^*$ and $P_V^{\perp} = I - VV^*$.

More specifically, in the SPSD or HPSD setting, for the least squares function $f(Z) = \frac{1}{2}\|Z - X\|_{\mathrm{F}}^2$, we have the following result.

Lemma 4.2.3 (Existence of gradient flow)**.** *Consider the manifold of SPSD or HPSD matrices $\mathcal{M}_r = \{Z \in \mathbb{S}_n \text{ or } \mathbb{H}_n,\ Z \succcurlyeq 0,\ \mathrm{rank}(Z) = r\}$. Consider the least squares loss function $f(Z) = \frac{1}{2}\|Z - X\|_{\mathrm{F}}^2$. Let $M := -\mathrm{grad} f(Z)$ denote its negative Riemannian gradient. Let $Z_0 \in \mathcal{M}_r$ be the initialization of the gradient flow at time $T = 0$, and $U_0 \in St(n, r)$, $S_0 \in \mathbb{S}_r$ nonsingular such that $Z_0 = U_0 S_0 U_0^\top$. Then there exists a unique gradient flow satisfying*

$$\begin{cases} \dot{U} = P_U^{\perp} MUS^{-1}, \\ \dot{S} = U^* MU, \end{cases}$$

for all $0 \leq T < \infty$.

Proof. The Riemannian gradient of the objective function $f(Z) = \frac{1}{2}\|Z - X\|_{\mathrm{F}}^2$ is $P_{T_Z}(Z - X)$. Plugging in $M = -P_{T_Z}(Z - X)$, and noticing that $P_U^{\perp} Z = 0$ and

[1] Strictly speaking, our ODE system is not an "approximation" but an exact characterization of the gradient flow. We stick to this terminology for ease of reference.

$P_U U = U$, we get the following ODE system:

$$\begin{cases} \dot{U} = P_U^{\perp} X U S^{-1}, \\ \dot{S} = -S + U^* X U. \end{cases}$$

It suffices to show that the ODE system does not blow up in finite time. We prove that for any $T_1 > 0$, $\sigma_{\min}(S)$ is bounded from below for all $T \in [0, T_1]$, where $\sigma_{\min}(S)$ is the smallest eigenvalue of $S \in S_r$.

At $T = 0$, we have $\sigma_{\min}(S) > 0$. At a given time T, let the multiplicity of $\sigma_{\min}(S)$ be j, i.e., $\sigma_{r-j}(S) > \sigma_{r-j+1}(S) = \ldots = \sigma_r(S)$. Denote $P_{U_{(r-j+1)\text{ to }r}}$ as the projection onto the corresponding eigen subspace. Using a similar argument as in [104], one can show that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{l=r-j+1}^{r} \sigma_l(S)\right) = \operatorname{tr}\left(P_{U_{(r-j+1)\text{ to }r}} \cdot \frac{\mathrm{d}}{\mathrm{d}t} S\right).$$

In particular, when $\sigma_r(S)$ is a simple eigenvalue and u_r is its eigenvector, this reduces to the classical result

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma_r(S) = u_r^* \left(\frac{\mathrm{d}}{\mathrm{d}t} S\right) u_r.$$

Note that $\frac{\mathrm{d}}{\mathrm{d}t} S = -S + U^* X U$ and X is positive semi-definite. Thus $\frac{\mathrm{d}}{\mathrm{d}t} S \succcurlyeq -S$, and we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{l=r-j+1}^{r} \sigma_l(S)\right) \geq \operatorname{tr}\left(P_{U_{(r-j+1)\text{ to }r}} \cdot (-S)\right) = -\sum_{l=r-j+1}^{r} \sigma_l(S).$$

In particular, when $\sigma_r(S)$ is a simple eigenvalue, one has

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma_r(S) \geq -\sigma_r(S).$$

By Grönwall's inequality, $\sigma_{\min}(S)$ decays no faster than exponentially fast. Thus it is bounded from below in any finite time interval. □

The dynamical low-rank approximation introduces a multiple-to-one mapping as a parameterization of $\mathcal{M}_r$. Let $\mathrm{St}(n, r)$ denote the n by r Stiefel manifold, i.e., $\mathrm{St}(n, r) = \{U \in \mathbb{F}^{n \times r} : U^* U = I_r\}$. Then we have that for S nonsingular,

$$\begin{array}{rcl} \mathrm{St}(n, r) \oplus \mathbb{S}_r & \to & \mathcal{M}_r \\ (U, S) & \mapsto & Z = U S U^*. \end{array}$$

Since S is not required to be diagonal, there are infinitely many tuples of (U, S) corresponding to the same Z, and these tuples are not equivalent under permutations. However, after we impose the constraint $\dot{U}^*U = 0$, from any initial tuple (U_0, S_0) there is a *unique* path in $\mathrm{St}(n, r) \oplus \mathbb{S}_r$ that describes the gradient flow of f according to Lemma 4.2.3. In other words, as long as the initial decomposition $Z_0 = U_0 S_0 U_0^*$ is given, the decomposition that satisfies the dynamical low-rank relation is uniquely determined along the whole trajectory.

The advantage of the dynamical low-rank approximation lies in the fact that the ODE system generically stays continuous. This is especially remarkable for the eigenvector matrix U. In comparison, SVD might enjoy uniqueness to some extent, but it is known to lose its differentiability when singular values coalesce [49], and that could only be fixed with the unrealistic assumption of analyticity [22].

Under the above parameterization, any isolated critical point $Z_\#$ on $\mathcal{M}_r$ corresponds to a critical set on $\mathrm{St}(n, r) \oplus \mathbb{S}_r$ consisting of infinitely many points, denoted by $\mathcal{N}_{Z_\#}$:

$$\mathcal{N}_{Z_\#} := \{(U_\#, S_\#) : U_\# S_\# U_\#^* = Z_\#\}.$$

Some constraints need to be imposed on the above decomposition to make it a valid parameterization for a spurious critical point. We will discuss it in more detail in Section 4.2.

In the following, we do not distinguish between the parameterized gradient flow on $\mathrm{St}(n, r) \oplus \mathbb{S}_r$ and the original gradient flow on $\mathcal{M}_r$ when there is no confusion. To prove the asymptotic escape of spurious critical points on $\mathcal{M}_r$, then, is to prove the asymptotic escape of spurious critical submanifolds on $\mathrm{St}(n, r) \oplus \mathbb{S}_r$.

Parameterization of $\overline{\mathcal{M}_r}$

As is mentioned in the previous subsection, using the parameterization $\mathrm{St}(n, r) \oplus \mathbb{S}_r \to \mathcal{M}_r$, each single critical point $Z_\#$ corresponds to a submanifold $\mathcal{N}_{Z_\#}$. In this subsection, we formally establish this result.

In order to use the dynamical low-rank approximation from Section 4.2, we decompose a rank-r matrix $Z \in \mathbb{S}_n$ into $Z = USU^*$, where $U \in \mathrm{St}(n, r)$ and

$S \in \mathbb{S}_r$. This decomposition differs from the eigenvalue decomposition in that S is not necessarily a diagonal matrix.

Consider a spurious critical point $Z_\# = U_1 D_1 U_1{}^* \in \mathcal{M}_s \subset \overline{\mathcal{M}_r} \backslash \mathcal{M}_r$, where $U_1 \in \mathbb{F}^{n \times s}$ represents the s eigenvectors that are also eigenvectors of X. We would like to determine a submanifold $\mathcal{N}_{Z_\#} \subset \mathrm{St}(n, r) \oplus \mathbb{S}_r$ that corresponds to $Z_\#$. Assume that

$$Z_\# = U_\# S_\# U_\#^*,$$

where

$$S_\# = P_\# \Sigma_\# P_\#^*$$

is the eigenvalue decomposition of $S_\#$. Then there exists $U_3 \perp U_1$, such that

$$U_\# = (U_1, U_3) P_\#^*, \qquad \Sigma_\# = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix}.$$

In addition, for $Z_\#$ to be a critical point of $f(Z) = \frac{1}{2}\|Z - X\|_{\mathrm{F}}^2$ on $\mathcal{M}_r$, we need $\operatorname{grad} f(Z_\#) = 0$, i.e., $P_{T_{Z_\#}}(Z_\# - X) = 0$. One can show that this gives

$$U_3 \perp U_X = (U_1, U_2).$$

In other words, U_3, the $n \times (r - s)$ matrix that makes up for the missing rank, should be chosen to be perpendicular to the missing component U_2. This also gives us $\lim_{Z \to Z_\#} \operatorname{grad} f(Z) = 0$, a property that will be useful in upcoming computation.

To sum up, a spurious critical point $Z_\# \in \mathcal{S}_\#$ can be parameterized as

$$\mathcal{N}_{Z_\#} = \left\{ (U_\#,\ S_\#) :\ U_\# = (U_1, U_3) P_\#^*,\ S_\# = P_\# \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix} P_\#^*,\ U_3 \perp U_X \right\},$$

where $P_\# \in \mathrm{SO}(r)$ is an orthonormal matrix.

Lemma 4.2.4. *$\mathcal{N}_{Z_\#}$ is an embedded submanifold of the manifold $\mathcal{M} := St(n, r) \oplus \mathbb{S}_r$.*

Proof. The proof of this lemma is very technical and is deferred to Section 4.6. The main idea is to invoke the definition of a submanifold directly and construct chart functions on $\mathcal{N}_{Z_\#}$. □

Rescaled gradient flow

Consider the dynamical low-rank description of the gradient flow in Lemma 4.2.3:

$$\begin{cases} \dot{U} = F(U,S) := P_U^{\perp} X U S^{-1}, \\ \dot{S} = H(U,S) := -S + U^* X U. \end{cases} \tag{DLRA}$$

The main tool for the proof of asymptotic escape is the following *rescaled gradient flow* ODE system:

$$\begin{cases} \dot{U} = \widetilde{F}(U,S) := P_U^{\perp} X U S^{-1} \cdot \sigma_{\min}(S), \\ \dot{S} = \widetilde{H}(U,S) := (-S + U^* X U) \cdot \sigma_{\min}(S). \end{cases} \tag{DLRA*}$$

Here $\sigma_{\min}(S)$ denotes the smallest eigenvalue of the $r \times r$ matrix S. In other words, the rescaled system (DLRA*) is just the original system (DLRA) times a scalar $\sigma_{\min}(S)$.

We first show that the rescaled system (DLRA*) is well-defined.

Lemma 4.2.5 (Continuity)**.** *The functions $\widetilde{F}(U,S)$ and $\widetilde{H}(U,S)$ are C^0 in $\mathcal{M}_r$.*

Proof. Inside $\mathcal{M}_r$, the matrix inverse S^{-1} is well-defined, so are the functions $F(U,S)$ and $H(U,S)$. Then use the fact that the smallest eigenvalue $\sigma_{\min}(S)$ is C^0 with respect to S. □

Lemma 4.2.6 (C^0-extension)**.** *The functions $\widetilde{F}(U,S)$ and $\widetilde{H}(U,S)$ can be extended continuously to $\mathcal{S}_{r-1}$.*

Proof. Take any $Z_\# \in \mathcal{S}_{r-1}$ with parameterization $Z_\# = U_\# S_\# U_\#^*$. It suffices to show that $\lim_{Z \to Z_\#} \widetilde{F}(U,S)$ and $\widetilde{H}(U,S)$ exist, and are independent of the specific choices of parameterization.

Let $S = P \Sigma P^*$ and $S_\# = P_\# \Sigma_\# P_\#^*$ be the eigenvalue decompositions of S and $S_\#$ respectively. Denote $p_i = P(:,i)$, and $p_{\#,i} = P_\#(:,i)$. Assume that $X = U_X D_X U_X^* = \sum_{i=1}^r d_i u_i u_i^*$. Since $Z_\# \in \mathcal{S}_{r-1}$, from the previous subsection, we know that

$$\Sigma_\# = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix},$$

where D_1 is an $(r-1)\times(r-1)$ diagonal matrix, $D_1 = \mathrm{diag}\{d_1,\ldots,d_{r-1}\}$. Moreover, when $\|S - S_\#\|_{\mathrm{F}} < \epsilon$ for small enough ϵ, by the sinΘ theorem (Lemma 2.4.1), we have

$$\Sigma = \mathrm{diag}\{\sigma_1,\ldots,\sigma_{r-1},\sigma_r\},$$

where

$$\begin{aligned}&\sigma_j > \min\{d_1.\ldots,d_{r-1}\} - \epsilon, \quad 1 \le j \le r-1;\\ &0 \le \sigma_r < \epsilon.\end{aligned}$$

In other words, σ_r and the rest of the eigenvalues of S are well-separated. Thus, when ϵ is small enough, we always have $\sigma_{\min}(S) = \sigma_r$.

Consider $\varphi(S) := S^{-1}\sigma_{\min}(S)$. When $\|S - S_\#\|_{\mathrm{F}} < \epsilon$, we have

$$\begin{aligned}\varphi(S) &= P\cdot\mathrm{diag}\{\sigma_1^{-1},\ldots,\sigma_{r-1}^{-1},\sigma_r^{-1}\}\cdot P^*\cdot\sigma_r\\ &= P\cdot\mathrm{diag}\left\{\frac{\sigma_r}{\sigma_1},\ldots,\frac{\sigma_r}{\sigma_{r-1}},1\right\}\cdot P^*\\ &= P\cdot\mathrm{diag}\left\{\frac{\sigma_r}{\sigma_1},\ldots,\frac{\sigma_r}{\sigma_{r-1}},0\right\}\cdot P^* + p_r p_r^*.\end{aligned}$$

Thus,

$$\begin{aligned}\lim_{S\to S_\#}\varphi(S) &= \lim_{S\to S_\#}\left(P\cdot\mathrm{diag}\left\{\frac{\sigma_r}{\sigma_1},\ldots,\frac{\sigma_r}{\sigma_{r-1}},0\right\}\cdot P^* + p_r p_r^*\right)\\ &= 0 \ + \ p_{\#,r}p_{\#,r}^*\\ &= p_{\#,r}p_{\#,r}^*.\end{aligned}$$

In other words, $\varphi(S)$ can be continuously extended to $S_\#$.

We can now compute the limits of $\widetilde{F}$ and $\widetilde{H}$. Note that

$$\widetilde{F}(U,S) = P_U^\perp XU\cdot\varphi(S).$$

Using the parameterization

$$Z_\# = U_\# S_\# U_\#^* : \quad U_\# = (U_1,U_3)P_\#^*, \quad S_\# = P_\#\begin{pmatrix}D_1 & 0\\ 0 & 0\end{pmatrix}P_\#^*,$$

we have

$$\begin{aligned}
\lim_{(U,S)\to(U_\#,S_\#)} \widetilde{F}(U,S) &= P_{U_\#}^{\perp} X U_\# \cdot \lim_{S\to S_\#} \varphi(S) \\
&= P_{U_\#}^{\perp} X U_\# \cdot p_{\#,r} p_{\#,r}^* \\
&= (I - P_{U_1} - P_{U_3}) \cdot (U_1 D_1 U_1^* + U_2 D_2 U_2^*) \cdot (U_1, U_3) P_\#^* \cdot p_{\#,r} p_{\#,r}^* \\
&= U_2 D_2 U_2^* \cdot (U_1, U_3) \cdot P_\#^* \cdot p_{\#,r} p_{\#,r}^* \\
&= 0.
\end{aligned}$$

As for $\widetilde{H}(U,S)$, since $H(U,S)$ is bounded and $\sigma_{\min}(S)$ converges to zero, we have

$$\lim_{(U,S)\to(U_\#,S_\#)} \widetilde{H}(U,S) = \lim_{(U,S)\to(U_\#,S_\#)} H(U,V,S) \cdot \sigma_{\min}(S) = 0.$$

Thus, $\widetilde{F}(U,S)$ and $\widetilde{H}(U,S)$ can both be extended continuously to $\mathcal{S}_{r-1}$, independent of the parameterization. □

Limit points of the rescaled system

In this subsection, we show that the ODE systems (DLRA) and (DLRA*) have the same limit points.

Lemma 4.2.7 (Existence of rescaled gradient flow)**.** *Consider the rescaled ODE system (DLRA*). Let $Z_0 \in \mathcal{M}_r$ be the initialization of the gradient flow at time $T = 0$, and $U_0 \in St(n,r)$, $S_0 \in \mathbb{S}_r$ nonsingular such that $Z_0 = U_0 S_0 U_0^\top$. Then there exists a unique gradient flow that satisfies (DLRA*) for all $T \in [0,\infty)$.*

Proof. The proof follows the same idea as that of Lemma 4.2.3. We show that within finite time, (U,S) remains in a region where $\widetilde{F}$ and $\widetilde{H}$ are Lipschitz continuous. Note that $\nabla_{S_{ij}}(S^{-1}) = -S^{-1} E_{ij} S^{-1}$ where E_{ij} is the indicator matrix of the (i,j)-entry. Note also that the smallest eigenvalue $\sigma_{\min}(S)$ is Lipschitz continuous with respect to S [84]. Thus the Lipschitz continuity of $\widetilde{F}$ and $\widetilde{H}$ holds if S^{-1} is bounded. This is true if $\sigma_{\min}(S)$ is bounded from below.

At a given time T, let the multiplicity of $\sigma_{\min}(S)$ be j, i.e., $\sigma_{r-j}(S) > \sigma_{r-j+1}(S) = \ldots = \sigma_r(S)$. Denote $P_{U_{(r-j+1)\text{ to }r}}$ as the projection onto the corresponding eigen subspace. Using a similar argument as in [104], we now have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{l=r-j+1}^{r}\sigma_l(S)\right) &= \operatorname{tr}\left(P_{U_{(r-j+1)\text{ to }r}}\cdot\frac{\mathrm{d}}{\mathrm{d}t}S\right)\\
&= \operatorname{tr}\left(P_{U_{(r-j+1)\text{ to }r}}\cdot(-S+U^*XU)\cdot\sigma_{\min}(S)\right)\\
&\geq \operatorname{tr}\left(P_{U_{(r-j+1)\text{ to }r}}\cdot(-S)\right)\cdot\sigma_{\min}(S)\\
&= -\left(\sum_{l=r-j+1}^{r}\sigma_l(S)\right)\cdot\sigma_{\min}(S).
\end{aligned}$$

In particular, when $\sigma_r(S)$ is a simple eigenvalue, this reduces to

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma_r(S) \geq -\sigma_r(S)^2.$$

Thus $\sigma_{\min}(S)$ decays no faster than geometrically due to Grönwall's inequality. Thus it is bounded from below in any finite time interval. □

Lemma 4.2.8 (Limit points). *Let $Z_0 \in \mathcal{M}_r$. Then the limit points of the ODE system (DLRA*) are the same as those of (DLRA). Moreover, the gradient flows starting from the same initial point always converge to the same limit point.*

Proof. Observe that the rescaled system (DLRA*) is just the original system (DLRA) multiplied by a scalar:

$$\begin{aligned}
\widetilde{F}(U,S) &= F(U,S)\cdot\sigma_{\min}(S),\\
\widetilde{H}(U,S) &= H(U,S)\cdot\sigma_{\min}(S).
\end{aligned}$$

Thus the gradient flow of the rescaled system follows the same path as the original system. In other words, let Z_t $(t \geq 0)$ and $\widetilde{Z}_t$ $(t \geq 0)$ be the solutions of (DLRA) and (DLRA*) starting from the same initial point Z_0, then for any time $t \geq 0$, there exists a corresponding time $w \geq 0$ such that $\widetilde{Z}_t = Z_w$.

When the time goes to infinity, both flows have limit points because both are minimizing flows of a coercive and lower-bounded function f. Denote them as $\widetilde{Z}_\infty$ and Z_∞ respectively. Then either $\widetilde{Z}_\infty = Z_\infty$, or there exists a finite T such that $\widetilde{Z}_\infty = Z_T$.

We now argue that only $\widetilde{Z}_\infty = Z_\infty$ is possible. Looking at the ODE system (DLRA*), a critical point has to satisfy either $F(U,S) = H(U,S) = 0$, or $\sigma_{\min}(S) = 0$. In the former case, $F(U,S) = H(U,S) = 0$ means such (U,S)

is stationary for (DLRA), so $\widetilde{Z}_\infty = Z_\infty$. In the latter case, such (U, S) has to be Z_∞ because we know that $\sigma_{\min}(S)$ cannot be zero at any finite time from Lemma 4.2.3. So either way, $\widetilde{Z}_\infty = Z_\infty$. Therefore, the limit points of (DLRA*) could only be those of (DLRA). □

By Lemma 4.2.8, if we can prove that gradient flows of (DLRA*) starting from random initializations almost surely avoid the spurious critical points in $\mathcal{S}_{r-1}$, we immediately have that the same results apply to (DLRA). In the next subsection, we will show that this is much easier to prove for the rescaled system than for the original system, because the points in $\mathcal{S}_{r-1}$ are now strict saddle points in the classical sense.

Landscape around the spurious critical points

We now analyze the landscape of (DLRA*) around the spurious critical points. In fact, we will show that the C^0-extension that we proved in Lemma 4.2.6 can be improved to a C^1-extension.

Lemma 4.2.9 (C^1-extension)**.** *Assume that the eigenvalues of the ground truth matrix X are all distinct. The functions $\widetilde{F}(U, S)$ and $\widetilde{H}(U, S)$ can be C^1-extended to $\mathcal{S}_{r-1}$.*

Proof. We first compute ∇F and ∇H in the interior of $\mathcal{M}_r$. In this region, S is non-singular, and all the derivatives are well defined. Let $\xi = (\xi_1, \xi_2)$ be a placeholder for the directional derivative, where ξ_1 and ξ_2 correspond to the direction of U and S respectively. Direct computation gives

$$\nabla F(U, S)[\xi] = \begin{pmatrix} -(U\xi_1^* + \xi_1 U^*)XUS^{-1} + P_U^\perp X\xi_1(S^{-1})^* \\ -P_U^\perp XUS^{-1}\xi_2 S^{-1} \end{pmatrix},$$

$$\nabla H(U, S)[\xi] = \begin{pmatrix} \xi_1^* XU + U^* X\xi_1 \\ -\xi_2 \end{pmatrix}.$$

To extend ∇F and ∇H themselves to $\mathcal{S}_{r-1}$ is impossible: S^{-1} is singular near $\mathcal{S}_{r-1}$, causing the derivatives to explode. We aim to show that it becomes possible with the rescaled system (DLRA*).

For this purpose, we define the following function, which is the directional derivative of $\varphi(S)$ along the direction η:

$$\psi(S, \eta) := \nabla\varphi(S)[\eta] = \nabla(S^{-1}\sigma_{\min}(S))[\eta].$$

We follow the same notations as before. Direct computation gives

$$\lim_{S\to S_\#}\psi(S,\eta)=\lim_{S\to S_\#}\left(-S^{-1}\eta S^{-1}\sigma_{\min}(S)+S^{-1}\cdot\nabla\sigma_{\min}(S)[\eta]\right).$$

We know from the proof of Lemma 4.2.6 that when $\|S-S_\#\|_{\mathrm{F}}<\epsilon$ for small enough ϵ, the larger eigenvalues σ_1 to σ_{r-1} and the smallest eigenvalue σ_r are well-separated. In fact, assuming that the eigenvalues of X are distinct, for small enough ϵ, all the eigenvalues of S are well-separated, and the corresponding eigenvectors are continuous with respect to the change of S. In this case, we know from [104] that

$$\nabla\sigma_r(S)[\eta]=p_r^*\eta p_r.$$

Thus, we have

$$\lim_{S\to S_\#}\psi(S,\eta)=\lim_{S\to S_\#}\left(-S^{-1}\eta S^{-1}\sigma_r+S^{-1}p_r^*\eta p_r\right).$$

For simplicity, we focus on the real case $\mathbb{F}=\mathbb{R}$. Since $\{p_ip_j^*\}_{i,j=1}^r$ form a complete orthogonal basis of $\mathbb{R}^{r\times r}$, we can write

$$\eta=\sum_{1\le i,j\le r}c_{ij}p_ip_j^*.$$

Such decomposition is continuous around $S_\#$, since all σ_i's are well-separated and all p_i's are continuous with respect to the change of S.

It now suffices to compute $\lim_{S\to S_\#}\psi(S,\eta)$ for $\eta=p_ip_j^*$, as $\psi(S,\eta)$ is linear in η. This comes in the following cases:

(1) If $i,\ j<r$:

$$\begin{aligned}\lim_{S\to S_\#}\psi(S,p_ip_j^*)&=\lim_{S\to S_\#}\left(-S^{-1}p_ip_j^*S^{-1}\sigma_r+S^{-1}p_r^*p_ip_j^*p_r\right)\\&=\lim_{S\to S_\#}\left(-P\Sigma^{-1}e_ie_j^*\Sigma^{-1}\sigma_rP^*+S^{-1}\cdot 0\right)\\&=\lim_{S\to S_\#}\left(-P\cdot 0\cdot P^*+0\right)\\&=0.\end{aligned}$$

(2) If $i<r,\ j=r$:

$$\begin{aligned}\lim_{S\to S_\#}\psi(S,p_ip_r^*)&=\lim_{S\to S_\#}\left(-S^{-1}p_ip_r^*S^{-1}\sigma_r+S^{-1}p_r^*p_ip_r^*p_r\right)\\&=\lim_{S\to S_\#}\left(-P\Sigma^{-1}e_ie_r^*\Sigma^{-1}\sigma_rP^*+S^{-1}\cdot 0\right)\\&=d_i^{-1}p_ip_r^*.\end{aligned}$$

(3) If $i = r$, $j < r$: Similar to the previous case,

$$\lim_{S\to S_\#} \psi(S, p_r p_j^*) = d_j^{-1} p_r p_j^*.$$

(4) If $i = j = r$:

$$\begin{aligned}
\lim_{S\to S_\#} \psi(S, p_r p_j^*) &= \lim_{S\to S_\#} \left(-S^{-1} p_r p_r^* S^{-1} \sigma_r + S^{-1} p_r^* p_r p_r^* v_r\right) \\
&= \lim_{S\to S_\#} \left(-P\Sigma^{-1} e_r e_r^* \Sigma^{-1} \sigma_r P^* + S^{-1}\right) \\
&= \lim_{S\to S_\#} \left(P \cdot \operatorname{diag}\left\{\sigma_1^{-1}, \ldots, \sigma_{r-1}^{-1}, -\sigma_r^{-1} + \sigma_r^{-1}\right\} \cdot P^*\right) \\
&= \lim_{S\to S_\#} \left(P \cdot \operatorname{diag}\left\{\sigma_1^{-1}, \ldots, \sigma_{r-1}^{-1}, 0\right\} \cdot P^*\right) \\
&= P_\# \cdot \operatorname{diag}\left\{\sigma_1^{-1}, \ldots, \sigma_{r-1}^{-1}, 0\right\} \cdot P_\#^*.
\end{aligned}$$

Therefore, $\psi(S, \eta)$ can be continuously extended to $S_\#$ for any η.

We now compute the derivatives of $\widetilde{F}$ and $\widetilde{H}$ at $Z_\#$. The directional derivative in U only involves $\varphi(S_\#)$, and we have

$$\begin{aligned}
\lim_{Z\to Z_\#} \nabla_U \widetilde{F}(U, S)[\xi_1] &= \lim_{Z\to Z_\#} \left(-(U\xi_1^* + \xi_1 U^*) XUS^{-1}\sigma_{\min}(S) + P_U^\perp X\xi_1 S^{-1}\sigma_{\min}(S)\right) \\
&= -(U_\#\xi_1^* + \xi_1(U_\#)^*) XU_\# \cdot \varphi(S_\#) + P_{U_\#}^\perp X\xi_1 \cdot \varphi(S_\#) \\
&= -(U_\#\xi_1^* + \xi_1(U_\#)^*) XU_\# \cdot p_r p_r^* + P_{U_\#}^\perp X\xi_1 p_r p_r^*.
\end{aligned}$$

As for the directional derivative in S, we now make use of $\psi(S_\#, \eta)$:

$$\begin{aligned}
\lim_{Z\to Z_\#} \nabla_S \widetilde{F}(U, S)[\xi_2] &= \lim_{Z\to Z_\#} \nabla_S \left(P_U^\perp XUS^{-1}\right)[\xi_2] \\
&= \lim_{Z\to Z_\#} \left(P_U^\perp XU\psi(S, \xi_2)\right) \\
&= P_{U_\#}^\perp XU_\# \cdot \psi(S_\#, \xi_2).
\end{aligned}$$

Since $P_{U_\#}^\perp XU_\# = 0$ and $\psi(S_\#, \xi_2)$ is bounded, we have

$$\lim_{Z\to Z_\#} \nabla_S \widetilde{F}(U, S)[\xi_2] = 0 \cdot \psi(S_\#, \xi_2) = 0.$$

Thus, the derivatives of $\widetilde{F}$ can be extended continuously to $Z_\#$, and we have

$$\lim_{Z\to Z_\#} \nabla \widetilde{F}(U, S) = \begin{pmatrix} -(U_\#\xi_1^* + \xi_1(U_\#)^*) XU_\# \cdot p_r p_r^* + P_{U_\#}^\perp X\xi_1 \cdot p_r p_r^* \\ 0 \end{pmatrix}.$$

As for the derivative of $\widetilde{H}$, we have

$$\lim_{Z \to Z_\#} \nabla \widetilde{H}(U,S)[\xi] = \lim_{Z \to Z_\#} \begin{pmatrix} (\xi_1^* XU + U^* X \xi_1) \cdot \sigma_{\min}(S) \\ -\xi_2 \cdot \sigma_{\min}(S) + (-S + U^* XU) \nabla_S \sigma_{\min}(S)[\xi_2] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Thus, we have shown that the derivatives of $\widetilde{F}(U,S)$ and $\widetilde{H}(U,S)$ can both be extended continuously to such $Z_\#$, which is equivalent to saying that the functions themselves can be C^1-extended to such $Z_\#$. □

The C^1-extension is crucial to the landscape analysis of the system (DLRA*) at the submanifolds corresponding to the rank-(r-1) spurious critical points. It enables us to compute the Jacobian right at those submanifolds, and determine its eigenvalues. We now show that those submanifolds are actually strict critical submanifolds of the system (DLRA*).

Lemma 4.2.10 (Strict critical submanifold)**.** *Assume that the eigenvalues of the ground truth matrix X are all distinct. Given a point $Z_\# \in \mathcal{S}_{r-1}$, let $\mathcal{N}_{Z_\#} = \{(U_\#, S_\#) : U_\# S_\# U_\#^* = Z_\#\}$ be the submanifold after parameterization that corresponds to $Z_\#$. Then $\mathcal{N}_{Z_\#}$ is a strict critical submanifold of the system (DLRA*) in the sense of Definition 3.3.3.*

Proof. The goal is to show that for any $(U_\#, S_\#) \in \mathcal{N}_{Z_\#}$, it is a hyperbolic point of the gradient flow with at least one escape direction, and all these points in $\mathcal{N}_{Z_\#}$ share a common escape direction perpendicular to the submanifold itself with a uniformly bounded eigenvalue. We will determine this escape direction by construction, using the results from the proof of Lemma 4.2.9.

Recall that $S = P\Sigma P^*$, $S_\# = P_\# \Sigma_\# P_\#^*$, and $X = U_X D_X U_X^* = \sum_{i=1}^r d_i u_i u_i^*$. Let

$$\xi = (\xi_1, \xi_2), \quad \xi_1 = u_r p_{\#,r}^*, \quad \xi_2 = 0.$$

Note that $X = U_1 D_1 U_1^* + U_2 D_2 U_2^*$, where $U_2 = u_r$, $D_2 = d_r$, and $U_2 D_2 U_2^*$ is the missing component in this spurious critical point $Z_\#$. In other words, we construct ξ exactly along the direction of this missing component. Using this property, we have that

$$\begin{aligned}
&\nabla_U \widetilde{F}(U,S)[\xi_1] \mid_{(U,S)=(U_\#,S_\#)} \\
&= -(U_\#\xi_1^* + \xi_1 U_\#^*)XU_\# \cdot p_{\#,r}p_{\#,r}^* + P_{U_\#}^\perp X\xi_1 \cdot p_{\#,r}p_{\#,r}^* \\
&= -(U_\# p_{\#,r}u_r^* + u_r p_{\#,r}^* U_\#^*)XU_\# \cdot p_{\#,r}p_{\#,r}^* + (I - P_{U_1}^\perp - P_{U_3}^\perp)X \cdot u_r p_{\#,r}^* \cdot p_{\#,r}p_{\#,r}^* \\
&= -\left((0,U_3)u_r^* + u_r(0,U_3)^*\right)\left(U_1D_1U_1^* + U_2D_2U_2^*\right) \cdot U_\# p_{\#,r}p_{\#,r}^* + U_2D_2U_2^* \cdot u_r p_{\#,r}^* \\
&= 0 + d_r u_r u_r^* \cdot u_r p_{\#,r}^* \\
&= d_r u_r p_{\#,r}^*,
\end{aligned}$$

and

$$\nabla_S \widetilde{F}(U,S)[\xi_2] \mid_{(U,S)=(U_\#,S_\#)}= 0.$$

Thus,

$$\nabla \widetilde{F}(U,S)[\xi] \mid_{(U,S)=(U_\#,S_\#)}= d_r u_r p_{\#,r}^* + 0 = d_r u_r p_{\#,r}^*.$$

Meanwhile,

$$\nabla \widetilde{H}(U,S)[\xi] \mid_{(U,S)=(U_\#,S_\#)}= 0.$$

Putting everything together, we have

$$\begin{aligned}
\nabla(\widetilde{F},\widetilde{H})[\xi] &= d_r \cdot (u_r p_{\#,r}^*, 0) \\
&= d_r \cdot \xi.
\end{aligned}$$

This means that $\xi = (u_r p_{\#,r}^*,\ 0)$ is an eigenvector of the Jacobian $\nabla(\widetilde{F},\widetilde{H})$ with eigenvalue d_r, which is positive.

Thus, for every tuple $(U_\#, S_\#)$ in $\mathcal{N}_{Z_\#}$, we have found an escape direction with uniform eigenvalue. So $\mathcal{N}_{Z_\#}$ is a strict critical submanifold as desired. □

Proof of the main result

We now prove Theorem 4.2.1 using the results from previous subsections.

Proof of Theorem 4.2.1. By Lemma 2.3.1, there are only finitely many spurious critical points that belong to $\mathcal{S}_{r-1}$. By Lemma 4.2.10, for each $Z_\# \in \mathcal{S}_{r-1}$, in the parameterized domain $\mathrm{St}(n,r) \oplus \mathbb{S}_r$, the corresponding submanifold $\mathcal{N}_{Z_\#}$ is a strict critical submanifold for the rescaled gradient flow. Since there are only finitely many of them, we can apply Theorem 3.2.1. This implies that the rescaled gradient flow in the parameterized domain almost never converges to $\cup_{Z_\# \in \mathcal{S}_{r-1}} \mathcal{N}_{Z_\#}$. Thus the rescaled gradient flow in the original domain $\mathcal{M}_r$ also almost never converges to $\mathcal{S}_{r-1}$. By Lemma 4.2.8, the original gradient flow has the same limit as the rescaled gradient flow. Thus the original gradient flow enjoys the same result, i.e., Prob $(\lim_{t\to\infty} Z_t \in S_{r-1}) = 0$. □

4.3 Main result for the gradient descent

The previous section has focused on the gradient flow. In this section we derive the result for the gradient descent, namely the asymptotic escape of the Riemannian gradient descent algorithm from the spurious critical points in $\mathcal{S}_{r-1}$.

Theorem 4.3.1 (Asymptotic escape of $\mathcal{S}_{r-1}$: gradient descent)**.** *Consider $f(Z) = \frac{1}{2}\|Z - X\|_{\mathrm{F}}^2$ where $X \in \mathcal{M}_r$ has distinct eigenvalues. Let $Z_0 \in \mathcal{M}_r$ be a random initialization, and $\{Z_k\}_{k=0}^{\infty}$ be the sequence generated by the following Riemannian gradient descent algorithm with varying step size:*

$$Z_{k+1} = R\left(Z_k - \alpha \cdot \sigma_r(Z_k) \cdot P_{T_{Z_k}}(\nabla f(Z_k))\right), \tag{4.2}$$

i.e., $\alpha_k = \alpha \cdot \sigma_r(Z_k)$, where $\sigma_r(Z_k)$ is the r-th eigenvalue of Z_k, and $\alpha > 0$. Assume that $Z_k \in \mathcal{M}_r$ for any $k < +\infty$, i.e., the sequence stays inside $\mathcal{M}_r$ at any finite step. Then we have

$$Prob\ (\lim_{k\to\infty} Z_k \in S_{r-1}) = 0.$$

In particular, this holds true for arbitrarily large $\alpha > 0$.

Remark 4.3.2. A few remarks are in order.

(1) The step size $\alpha_k = \alpha \cdot \sigma_r(Z_k)$ is varying but not necessarily diminishing. As long as the sequence eventually escapes the spurious critical points, $\sigma_r(Z_k)$ converges to $\sigma_r(X)$ and $\alpha_k = \alpha \cdot \sigma_r(Z_k)$ does not diminish. See Figure 4.2b for a numerical illustration.

(2) The reason for the choice $\alpha_k = \alpha \cdot \sigma_r(Z_k)$ is similar to the rescaling of the ODE system (DLRA*) in the previous section. Namely, this makes the Jacobian of the iteration function C^1-extendable to the rank-(r-1) spurious critical points in $\mathcal{S}_{r-1}$, using the same techniques as in the proof of Lemma 4.2.9.

Proof of Theorem 4.3.1. We use the same notations as before, namely $Z = USU^*$, $S = P\Sigma P^*$, $S_\# = P_\# \Sigma_\# P_\#^*$, and $X = U_X D_X U_X^* = \sum_{i=1}^r d_i u_i u_i^*$. We also let $Z = U_Z \Sigma U_Z^*$ denote the eigen decomposition of Z, which implies $U_Z = U \cdot P^*$. We let $\widetilde{U} \in \mathrm{St}(n, n-r)$ be the orthogonal complement of U. It is also the orthogonal complement of U_Z. Since $U = (U_1, U_3)$, where $U_3 \perp U_2$, we know that $\mathrm{span}\{U_2\} \subset \mathrm{span}\{\widetilde{U}\}$. Without loss of generality, we let U_2 be the first column of $\widetilde{U}$.

Consider the iteration function

$$\begin{aligned}\phi(Z) &= R\left(Z - \alpha \cdot \sigma_r(Z) \cdot P_{T_Z}(\nabla f(Z))\right) \\ &= R\left(Z - \alpha \cdot \sigma_r(Z) \cdot \mathrm{grad} f(Z)\right).\end{aligned} \tag{4.3}$$

Here $\mathrm{grad} f(Z)$ is the Riemannian gradient. The Jacobian of the iteration function is

$$D\phi(Z) = I - \alpha \cdot \left(\sigma_r(Z) \cdot \mathrm{Hess} f(Z) + D\sigma_r(Z) \cdot \mathrm{grad} f(Z)\right).$$

It has been shown in [125] that

$$\mathrm{Hess} f(Z)[\xi] = \xi + P_{U_Z}^{\perp}(Z - X)\widetilde{U} N \Sigma^{-1} U_Z^* + U_Z \Sigma^{-1} N^* \widetilde{U}^* (Z - X) P_{U_Z}^{\perp}, \tag{4.4}$$

where the vector ξ is parameterized as

$$\xi = U_Z M U_Z^* + U_Z N \widetilde{U}^* + \widetilde{U} N^* U_Z^*, \quad M \in \mathbb{F}^{r \times r}, \quad N \in \mathbb{F}^{r \times (n-r)}.$$

In particular, when $\mathbb{F} = \mathbb{R}$, the degree of freedom of ξ is $\frac{r(2n-r+1)}{2}$. It is equal to the dimension of the tangent space that ξ lies in, which is the same as the dimension of the manifold.

Consider $\lim_{Z \to Z_\#} D\phi(Z)$ for $Z_\# \in \mathcal{S}_{r-1}$. Note that the parameterization from Section 4.2 ensures that $\mathrm{span}\{U_3\} \perp \mathrm{span}\{U_1, U_2\}$, so that $Z_\#$ is a valid critical point, i.e., $\mathrm{grad} f(Z_\#) = 0$. Plugging Equation (4.4) into Equation (4.3), we have

$$\begin{aligned}
D\phi(Z_\#)[\xi] &:= \lim_{Z\to Z_\#} D\phi(Z)[\xi] \\
&= \xi - \alpha \cdot \left(\lim_{Z\to Z_\#} (\sigma_r(Z) \cdot \mathrm{Hess} f(Z)[\xi]) + D\sigma_r(Z)[\xi] \cdot \mathrm{grad} f(Z_\#) \right) \\
&= \xi - \alpha \cdot \left(\lim_{Z\to Z_\#} (\sigma_r(Z) \cdot \mathrm{Hess} f(Z)[\xi]) \right) \\
&= \xi - \alpha \cdot \left(\lim_{Z\to Z_\#} \left(\sigma_r(Z)\cdot\xi - P_U^\perp (Z-X)\widetilde{U}N\Sigma^{-1}U^* - U\Sigma^{-1}N^*\widetilde{U}^*(Z-X)P_U^\perp \right) \right) \\
&= \xi - \alpha \cdot \left(0\cdot\xi - U_2D_2{U_2}^\top \widetilde{U}N_2 \left(\lim_{\Sigma\to\Sigma_\#} \Sigma^{-1}\sigma_r \right) U_\#^* - U_\# \left(\lim_{\Sigma\to\Sigma_\#} \Sigma^{-1}\sigma_r \right) N^*\widetilde{U}^*U_2D_2{U_2}^\top \right) \\
&= \xi + \alpha \cdot \left(U_2D_2{U_2}^\top \widetilde{U}N_2 \left(\lim_{\Sigma\to\Sigma_\#} \Sigma^{-1}\sigma_r \right) U_\#^* + U_\# \left(\lim_{\Sigma\to\Sigma_\#} \Sigma^{-1}\sigma_r \right) N^*\widetilde{U}^*U_2D_2{U_2}^\top \right).
\end{aligned}$$

Here, similar to the proof of Lemma 4.2.6, we have

$$\lim_{\Sigma\to\Sigma_\#} \Sigma^{-1}\sigma_r = \mathrm{diag}\{0,\ldots,0,1\} = e_r e_r^*.$$

Thus, it follows that

$$D\phi(Z_\#)[\xi] = \xi + \alpha \cdot \left(U_2D_2{U_2}^\top \widetilde{U}Ne_re_r^*U_\#^* + U_\#e_re_r^*N^*\widetilde{U}^*U_2D_2{U_2}^\top \right).$$

Note that without loss of generality, we have let U_2 be the first column of $\widetilde{U}$. Thus we have

$$\begin{aligned}
D\phi(Z_\#)[\xi] &= \xi + \alpha \cdot \left(U_2D_2(1,0,\ldots,0)(Ne_r)\begin{pmatrix} {U_2}^\top \\ 0 \\ \vdots \\ 0 \end{pmatrix} + (U_2,\, 0,\ldots,0)(e_r^*N^*)\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} D_2{U_2}^\top \right) \\
&= \xi + \alpha \cdot 2N(1,1)\cdot U_2D_2{U_2}^\top.
\end{aligned}$$

We can immediately read the eigenvalues and eigenvectors of $D\phi(Z_\#)$ from the above expression. Specifically, when $\mathbb{F}=\mathbb{R}$, $D\phi(Z_\#)$ has

(1) One eigenvector $\xi = UN\widetilde{U}^* + \widetilde{U}NU^*$ with $N = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix}$, whose corresponding eigenvalue is $\lambda = 1 + 2\alpha\cdot D_2 > 1$;

(2) $(\frac{r(2n-r+1)}{2} - 1)$ eigenvectors with eigenvalues $\lambda = 1$.

The complex case $\mathbb{F} = \mathbb{C}$ is similar except that the dimensionality is different.

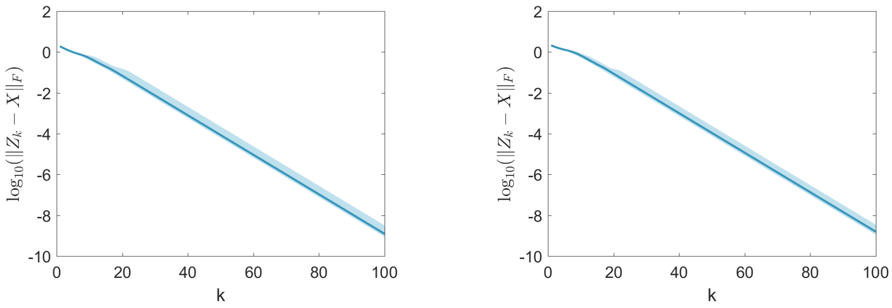
Now that $D\phi(Z_\#)$ has one eigenvalue greater than 1, while the rest of the eigenvalues are equal to 1. By [120, Theorem III.7], there is an unstable manifold and a center manifold in the neighborhood of $Z_\#$, which can be extended globally. The existence of the unstable manifold ensures that $Z_\#$ is an asymptotic unstable fixed point of the iteration function $\phi(Z_\#)$. Thus, the Riemannian gradient descent algorithm with varying step size (4.2) almost surely escapes $\mathcal{S}_{r-1}$.

In particular, $D\phi(Z_\#)$ is always a local diffeomorphism independent of the choice of α, as its only eigenvalues are 1 and $1 + \alpha D_2$. Therefore, the result of Theorem 4.3.1 holds true for arbitrarily large $\alpha > 0$. □

4.4 Numerical experiments

In this section, we present some numerical experiments to illustrate our theoretical results in Theorem 4.2.1 and 4.3.1. We also provide some evidence in support of conjectures beyond the previous theorem and lemma.

In all experiments, we consider the real SPSD case $\mathbb{F} = \mathbb{R}$, and let $n = 100$, $r = 5$, and we use the same ground truth matrix $X \in \mathcal{M}_r$ with distinct singular values. We use the Riemannian gradient descent algorithm to minimize $f(Z) = \frac{1}{2}\|Z - X\|_{\mathrm{F}}^2$. The experiments only differ by the sampling rule and the choice of the step sizes α_k. Each figure is generated by repeating the experiment 100 times. The shaded area represents the range of the data and the solid line represents the median.

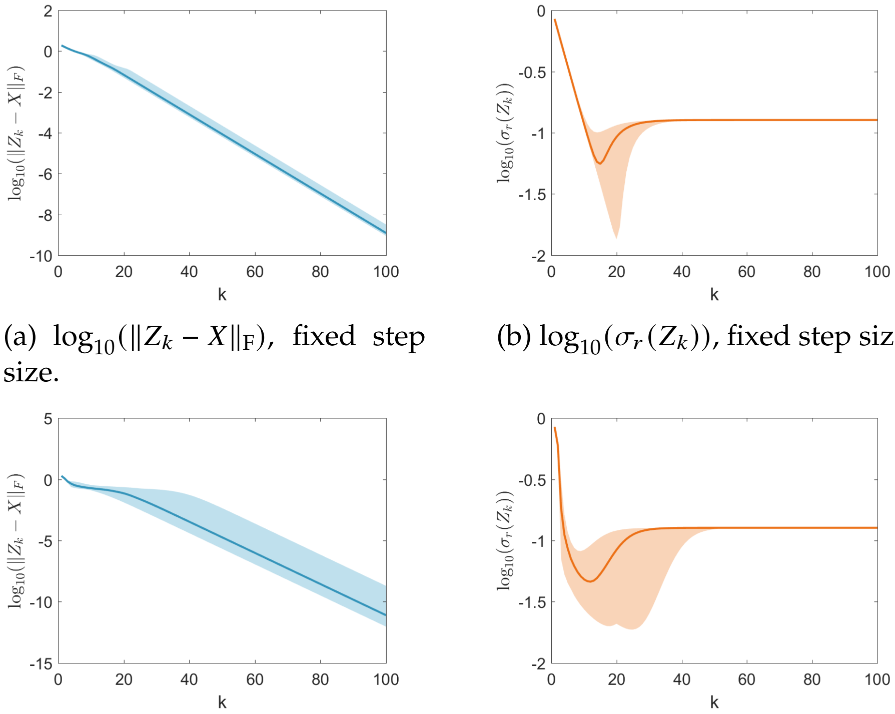


(a) Local escape near $\mathcal{S}_{r-1}$. (b) Local escape near $\mathcal{S}_{r-2}$.

Figure 4.1: Escape of spurious critical points.

The first experiment is performed near a rank-(r-1) spurious critical point $Z_\#^{(1)} \in \mathcal{S}_{r-1}$. The initial points are randomly sampled in the local neighborhood of $Z_{\#}^{(1)}$. The step size is fixed to be $\alpha_k \equiv \alpha = 0.2$. Figure 4.1a shows the log10 distance between Z_k and X. It can be seen that in all the repeated experiments, the sequence always succeeds to escape $Z_{\#}^{(1)}$ and converge to X.

To verify whether $\mathcal{S}_s$ ($s < r - 1$) incurs the same behavior, we repeat the experiment with $Z_{\#}^{(2)} \in \mathcal{S}_{r-2}$. It can be seen in Figure 4.1b that the phenomenon is indeed the same. Thus we conjecture that a similar result as Theorem 4.2.1 holds for those $\mathcal{S}_s$ with $s < r - 1$ as well. Proof of such result is left for future work.



(a) $\log_{10}(\|Z_k - X\|_{\mathrm{F}})$, fixed step size.

(b) $\log_{10}(\sigma_r(Z_k))$, fixed step size.

(c) $\log_{10}(\|Z_k - X\|_{\mathrm{F}})$, varying step size.

(d) $\log_{10}(\sigma_r(Z_k))$, varying step size.

Figure 4.2: Comparison of fixed and varying step sizes.

Next, we investigate Theorem 4.3.1 and the varying step size $\alpha_k = \alpha \cdot \sigma_r(Z_k)$. Figures 4.2a and 4.2b are the results with a fixed step size $\alpha_k \equiv 0.2$. Figures 4.2c and 4.2d are the results with varying step sizes $\alpha_k = 2\sigma_r(Z_k)$. The left are the distances to the ground truth X. The right are the log values of $\sigma_r(Z_k)$ along the iterative path. We can see that first of all, the iterative sequences always escape all spurious critical points and converge to the ground truth. Moreover, the value of $\sigma_r(Z_k)$ is never too small, but soon converges to the smallest eigenvalue of X. This helps illustrate that the varying step size $\alpha_k = \alpha \cdot \sigma_r(Z_k)$ is *not a diminishing step size* in practice, but is rather always above a certain value.

4.5 Discussion

In this chapter, we discuss the asymptotic escape of the spurious critical points on the low-rank matrix manifold. The goal is to shed some light on the nonclosedness of the low-rank matrix manifold $\mathcal{M}_r$ and justify the global use of Riemannian gradient descent on the manifold. To this end, we first point out the existence of a set of spurious critical points $\mathcal{S}_\# \subset \overline{\mathcal{M}_r} \backslash \mathcal{M}_r$ and discuss its singularity. We then use a rescaled gradient flow combined with the dynamical low-rank approximation to describe the local landscape, which enables us to eliminate the singularity and prove the asymptotic escape result. We also present a corresponding result for the gradient descent. Numerical experiments are provided to illustrate the theoretical results.

Though this study is focused on $\mathcal{S}_{r-1}$, the asymptotic escape is empirically observed for $\mathcal{S}_s$ with $s \leq r-2$ as well. In fact, all spurious critical points in $\mathcal{S}_\#$ are observed to be asymptotically unstable in practice, which can be seen from the numerical experiments. The current rescaled gradient flow (DLRA*) loses both C^0- and C^1-extensions at $\mathcal{S}_s$ with $s \leq r-2$. This is because the continuity of eigenvalues and eigenvectors are only possible when only one of the eigenvalues is approaching zero. Extension of the result to the case $s \leq r-2$ is left for future work. On the other hand, the assumption that the eigenvalues of X are distinct is not an essential assumption, and can easily be removed.

Even though the result for the gradient descent calls for a step size $\alpha \cdot \sigma_r(Z)$, this is not a diminishing step size. As long as the sequence eventually escapes the apocalyptic points, $\sigma_r(Z)$ converges to $\sigma_r(X)$ and α_k does not diminish. This is supported by numerical observations.

In addition to the asymptotic result in this chapter, a non-asymptotic result on the number of steps needed to escape the spurious critical points can be found in Chapter 5. There it is shown that the converging set of the spurious critical points can be upper bounded by a small positive measure. With high probability, one has nearly linear convergence rate toward the ground truth. The two sides of the story complement each other and provide a complete picture of the unique structure of the low-rank matrix manifold.

4.6 Proof of Lemma 4.2.4

We recall the previous lemma below.

Lemma 4.2.4. Define

$$\mathcal{N}_{Z_\#} := \left\{ (U_\#,\, S_\#) : \; U_\# = (U_1, U_3)P_\#^*, \;\; S_\# = P_\# \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix} P_\#^*, \; U_3 \perp U_X \right\}.$$

Then $\mathcal{N}_{Z_\#}$ is an embedded submanifold of the manifold $\mathcal{M} := \mathrm{St}(n,r) \oplus \mathbb{S}_r$.

The intuition behind Lemma 4.2.4 is that the set $\mathcal{N}_{Z_\#}$ is a subset of $\mathcal{M}$ characterized by some algebraic constraints, namely $U_\# S_\# U_\#^* = Z_\#$ and $U_3 \perp U_X$. As is often the case, one would expect such algebraic constraints to give an embedded submanifold. We will make this intuition rigorous in this section.

We note that traditionally, embedded submanifold is proved by the submersion theorem, i.e., by showing that the set is the preimage of a regular value of a submersive mapping. But this approach does not work here because $Z_\#$ is not a regular value. Instead, we need to go back to the definition of a submanifold and construct chart functions on $\mathcal{N}_{Z_\#}$ directly.

Below are some auxiliary results from the literature.

Lemma 4.6.1 ([1, Proposition 3.3.2]). *A subset $\mathcal{N}$ of a manifold $\mathcal{M}$ is a d-dimensional embedded submanifold of $\mathcal{M}$ if and only if, around each point $x \in \mathcal{N}$, there exists a chart $(\mathcal{U}, \varphi)$ of $\mathcal{M}$ such that $\mathcal{N} \cap \mathcal{U}$ is a φ-coordinate slice of $\mathcal{U}$, i.e.,*

$$\mathcal{N} \cap \mathcal{U} = \{x \in \mathcal{U} : \; \varphi(x) \in \mathbb{R}^d \times \mathbf{0}\}.$$

In this case, the chart $(\mathcal{N} \cap \mathcal{U}, \varphi)$, where φ is seen as a mapping into $\mathbb{R}^d$, is a chart of the embedded submanifold $\mathcal{N}$.

By Lemma 4.6.1, if we can construct an atlas of $\mathcal{M}$ and an atlas of $\mathcal{N}_{Z_\#}$, such that the charts in the latter atlas are coordinate slices of the charts in the former atlas, then $\mathcal{N}_{Z_\#}$ is an embedded submanifold of $\mathcal{M}$. This approach is less common than the traditional submersion theorem approach, but is necessary for our problem.

Lemma 4.6.2 ([57]). *For the real Stiefel manifold $St(n,k)$, there exists an atlas $\cup_Q(\mathcal{U}_Q, \varphi_Q)$ of the Stiefel manifold. Namely, for each chart $(\mathcal{U}_Q, \varphi_Q)$, Q is a matrix in $St(n,k)$, and the function φ_Q can be expressed as*

$$\begin{aligned} \varphi_Q: \quad & \mathcal{U}_Q \to Skew(k) \oplus \mathbb{R}^{(n-k)\times k}, \\ & U \mapsto (\Omega_{11}, \Omega_{21}), \end{aligned}$$

where

$$\begin{aligned} &\Omega_{11} = (U_1^\top + Q_1^\top)^{-1} \left(Q_1^\top U_1 + U_2^\top Q_2 - U_1^\top Q_1 - Q_2^\top U_2\right) (U_1 + Q_1)^{-1}, \quad \Omega_{11} = -\Omega_{11}^\top, \\ &\Omega_{21} = (U_2 - Q_2)(U_1 + Q_1)^{-1}, \end{aligned}$$

and $U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$, $Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$ are the block forms of U and Q respectively. Such a chart function is defined on the subset $\mathcal{U}_Q \subset St(n,k)$ which covers all of the manifold $St(n,k)$ except a zero-measure set.

In particular, if $Q = \begin{pmatrix} I_k \\ \mathbf{0} \end{pmatrix}$, then

$$\begin{aligned} \Omega_{11} &= (U_1^\top + I_k)^{-1} \left(U_1 - U_1^\top\right) (U_1 + I_k)^{-1}, \\ \Omega_{21} &= U_2 (U_1 + I_k)^{-1}. \end{aligned}$$

Lemma 4.6.2 provides a neat construction of charts on the Stiefel manifold. In fact, we only need two charts to cover the whole manifold, if we choose any two Q's that do not share any left singular vector. We will use this construction frequently in the proof of Lemma 4.2.4.

We are now ready to prove Lemma 4.2.4.

Proof of Lemma 4.2.4. We restrict our attention to the real case $\mathbb{F} = \mathbb{R}$. The complex case $\mathbb{F} = \mathbb{C}$ is very similar except that the dimensionalities of some manifolds in the subsequent proof are slightly different.

We aim to construct explicit charts of $\mathcal{M} = \mathrm{St}(n,r) \oplus \mathbb{S}_r$, and explicit charts of $\mathcal{N}_{Z_\#}$, such that the latter are the coordinate slices of the former. For clarity, we will first write out the charts of $\mathcal{N}_{Z_\#}$, and then express them as coordinate slices of charts of $\mathcal{M}$.

Step 1: Construct charts of $\mathcal{N}_{Z_\#}$.

For any $(U, S) \in \mathcal{N}_{Z_\#}$, we rewrite U and S as the following:

$$\begin{aligned} &U = U_1 P_1^\top + U_3 P_2^\top, \qquad S = P_1 D_1 P_1^\top, \\ &\text{where} \quad P_1 \in \mathbb{R}^{r\times s}, \quad P_2 \in \mathbb{R}^{r\times(r-s)}, \quad P = (P_1, P_2) \in \mathrm{SO}(r). \end{aligned}$$

We argue that there exists a mapping from every P_1 to a unique P_2. An intuitive explanation is that P_2 can always be uniquely determined by a Gram-Schmidt process starting from the identity matrix. Thus we can write $P_2 = \mathcal{P}_2(P_1)$ where $\mathcal{P}_2 : \mathbb{R}^{r\times s} \to \mathbb{R}^{r\times(r-s)}$ is a function. Therefore, any $(U, S) \in \mathcal{N}_{Z_\#}$ can be re-parameterized using only (P_1, U_3). We write this re-parameterization as a function f:

$$\begin{aligned} f: \quad \mathcal{N}_{Z_\#} &\to \mathrm{St}(r, s) \oplus \widetilde{\mathrm{St}}(n, r-s; U_X); \\ (U, S) &\mapsto (P_1, U_3). \end{aligned}$$

Here $\mathrm{St}(r, s)$ is a Stiefel manifold, and $\widetilde{\mathrm{St}}(n, r-s; U_X)$ is a constrained Stiefel manifold:

$$\widetilde{\mathrm{St}}(n, r-s; U_X) := \{U_3 : \; U_3 \in \mathrm{St}(n, r-s), \; U_3 \perp U_X\}, \quad \text{where } U_X = (U_1, U_2).$$

We now construct charts for P_1 and U_3 respectively. The domain of P_1 is the Stiefel manifold $\mathrm{St}(r, s)$. By Lemma 4.6.2, there exists an atlas where every chart function maps to $\mathrm{Skew}(s) \oplus \mathbb{R}^{(r-s)\times s}$. Let $g^{(1)}$ be one such chart function:

$$\begin{aligned} g^{(1)}: \quad \mathrm{St}(r, s) &\to \mathrm{Skew}(s) \oplus \mathbb{R}^{(r-s)\times s} \\ P_1 &\mapsto (\Omega_{11}, \Omega_{21}). \end{aligned}$$

The domain of U_3 is the constrained Stiefel manifold $\widetilde{\mathrm{St}}(n, r-s; U_X)$. Here $U_X \in \mathrm{St}(n, r)$ is the eigenvectors matrix of the ground truth X, which is fixed. To construct a chart of $\widetilde{\mathrm{St}}(n, r-s; U_X)$, we first construct a mapping $g^{(2)}$ according to Lemma 4.6.2, such that

$$\begin{aligned} g^{(2)}: \quad \widetilde{\mathrm{St}}(n, r-s; U_X) &\to \mathrm{Skew}(r-s) \oplus \widetilde{\mathbb{R}}^{(n-(r-s))\times(r-s)}, \\ U_3 &\mapsto (\Lambda_{11}, \Lambda_{21}). \end{aligned}$$

The domain of Λ_{21} is $\widetilde{\mathbb{R}}^{(n-(r-s))\times(r-s)}$, which is a constrained set. To express the constraints $U_3 \perp U_X$ in terms of constraints on Λ_{21}, we write $U_3 = \begin{pmatrix} U_{3,1} \\ U_{3,2} \end{pmatrix}$

and $U_X = \begin{pmatrix} U_{X,1} \\ U_{X,2} \end{pmatrix}$. Assume without loss of generality that $g^{(2)}$ is constructed by picking $Q = (I_{r-s}, \mathbf{0})^\top$ in Lemma 4.6.2. Then

$$\Lambda_{21} = U_{3,2}(U_{3,1} + I_{r-s})^{-1}.$$

Since $U_3 \perp U_X$, we have

$$U_X^\top U_3 = U_{X,1}^\top U_{3,1} + U_{X,2}^\top U_{3,2} = 0.$$

Thus,

$$U_{X,2}^\top U_{3,2} = -U_{X,1}^\top U_{3,1}.$$

This gives us

$$U_{X,2}^\top \Lambda_{21} = -U_{X,1}^\top U_{3,1}(U_{3,1} + I_{r-s})^{-1}.$$

These are linear constraints on Λ_{21}.

Let g be the concatenation of $g^{(1)}$ and $g^{(2)}$, then we have a re-parameterization of (P_1, U_3) as follows:

$$\begin{aligned} g: \quad \mathrm{St}(r,s) \oplus \widetilde{\mathrm{St}}(n, r-s) &\to \mathrm{Skew}(s) \oplus \mathbb{R}^{(r-s)\times s} \oplus \mathrm{Skew}(r-s) \oplus \widetilde{\mathbb{R}}^{(n-(r-s))\times(r-s)}; \\ (P_1, U_3) &\mapsto (\Omega_{11}, \Omega_{21}, \Lambda_{11}, \Lambda_{21}). \end{aligned}$$

Here $\widetilde{\mathbb{R}}^{(n-(r-s))\times(r-s)}$ is the submanifold of $\mathbb{R}^{(n-(r-s))\times(r-s)}$ defined by the linear constraints that we derived:

$$\widetilde{\mathbb{R}}^{(n-(r-s))\times(r-s)} := \left\{\Lambda_{21} \in \mathbb{R}^{(n-(r-s))\times(r-s)} : \; U_{X,2}^\top \Lambda_{21} = -U_{X,1}^\top U_{3,1}(U_{3,1} + I_{r-s})^{-1}\right\}.$$

Let Λ_{21}° be an arbitrary solution to the equation $U_{X,2}^\top \Lambda_{21} = -U_{X,1}^\top U_{3,1}(U_{3,1} + I_{r-s})^{-1}$. Then

$$\widetilde{\mathbb{R}}^{(n-(r-s))\times(r-s)} = \Lambda_{21}^\circ + \mathrm{Ker}(U_{X,2}^\top).$$

By finding an orthogonal basis for $\mathrm{Ker}(U_{X,2}^\top)$, it is easy to construct a chart function

$$\begin{aligned} h: \quad \widetilde{\mathbb{R}}^{(n-(r-s))\times(r-s)} &\to \mathbb{R}^{(n-(r-s))(r-s)-r(r-s)} \\ \Lambda_{21} &\mapsto \Gamma. \end{aligned}$$

Putting everything together, we have that

$$\begin{aligned}\varphi := (\mathrm{id}, h) \circ g \circ f : \quad \mathcal{N}_{Z_\#} &\to \mathrm{Skew}(s) \oplus \mathbb{R}^{(r-s)\times s} \oplus \mathrm{Skew}(r-s) \oplus \mathbb{R}^{(n-(r-s))(r-s)-r(r-s)};\\ (U, S) &\mapsto (\Omega_{11}, \Omega_{21}, \Lambda_{11}, \Gamma).\end{aligned}$$

This is a chart function for the whole $\mathcal{N}_{Z_\#}$ except a zero-measure set. Varying $g^{(1)}$ and $g^{(2)}$ as needed, we have the atlas for the whole $\mathcal{N}_{Z_\#}$.

Step 2: Express the charts of $\mathcal{N}_{Z_\#}$ as coordinate slices of charts of $\mathcal{M}$.

To express things into coordinate slices, we will work the other way around: we extend the chart function φ into a chart function $\widetilde{\varphi}$ defined on $\mathcal{M} = \mathrm{St}(n, r) \oplus \mathbb{S}_r$.

For any $(U, S) \in \mathcal{M} = \mathrm{St}(n, r) \oplus \mathbb{S}_r$, we construct a re-parameterization as follows:

$$\begin{aligned}&U = \Big((U_1R_1, 0) + (M_4, U_3R_2)\Big)\begin{pmatrix}P_1^\top\\ P_2^\top\end{pmatrix}, \qquad S = (P_1, P_2)\widetilde{S}\begin{pmatrix}P_1^\top\\ P_2^\top\end{pmatrix},\\ &\text{where}\quad P_1 \in \mathbb{R}^{r\times s}, \quad P_2 \in \mathbb{R}^{r\times(r-s)}, \quad P = (P_1, P_2) \in \mathrm{SO}(r),\\ &U_3 \in \widetilde{\mathrm{St}}(n, r-s;\, U_1), \quad M_4 \in \widetilde{\mathbb{R}}^{n\times s}, \quad \widetilde{S} \in \mathbb{S}_r,\\ &R_1 \in \mathrm{upper}(s, s), \quad R_2 \in \widetilde{\mathrm{upper}}(r-s, r-s).\end{aligned}$$

The domain of P_1 is $\mathrm{St}(r, s)$. P_2 is still uniquely determined by P_1 as before. The domain of U_3 is the constrained Stiefel manifold $\widetilde{\mathrm{St}}(n, r-s;\, U_1) := \{U_3 :\ U_3 \in \mathrm{St}(n, r-s),\ U_3 \perp U_1\}$. Note that the constraints are only in terms of U_1 instead of $U_X = (U_1, U_2)$. The domain of M_4 is the linearly constrained subspace $\widetilde{\mathbb{R}}^{n\times s} := \{M_4 \in \mathbb{R}^{n\times s},\ M_4 \perp U_1\}$. The domain of $\widetilde{S}$ is $\mathbb{S}_r$. The domain of R_1 is the subspace of $s \times s$ upper triangular matrices. The domain of R_2 is the subspace of $(r-s) \times (r-s)$ upper triangular matrices, but with some constraints that will be specified later. We define the following mapping:

$$\begin{aligned}\widetilde{f} : \quad \mathrm{St}(n, r) \oplus \mathbb{S}_r &\to \mathrm{St}(r, s) \oplus \widetilde{\mathrm{St}}(n, r-s;\, U_1) \oplus \widetilde{\mathbb{R}}(n, s) \oplus \widetilde{\mathrm{upper}}(r-s, r-s) \oplus \mathbb{S}_r \oplus \mathrm{upper}(s, s);\\ (U, S) &\mapsto \left(P_1, U_3, M_4, R_2 - I_{r-s}, \widetilde{S} - \begin{pmatrix}D_1 & 0\\ 0 & 0\end{pmatrix}, R_1 - I_s\right).\end{aligned}$$

The mapping $\widetilde{f}$ is written in such a way because, if $(U, S) \in \mathcal{N}_{Z_\#}$, then the last few components are all zero, and f is just a coordinate slice of $\widetilde{f}$:

$$\widetilde{f}(U, S) = (P_1, U_3, 0, 0, 0, 0)\,.$$

For the first part of the image of $\widetilde{f}$, we apply g as before:

$$g: \quad \mathrm{St}(r,s) \oplus \widetilde{\mathrm{St}}(n, r-s) \to \mathrm{Skew}(s) \oplus \mathbb{R}^{(r-s)\times s} \oplus \mathrm{Skew}(r-s) \oplus \widehat{\mathbb{R}}^{(n-(r-s))\times(r-s)};$$
$$(P_1, U_3) \mapsto (\Omega_{11}, \Omega_{21}, \Lambda_{11}, \Lambda_{21}).$$

However, the set $\widehat{\mathbb{R}}^{(n-(r-s))\times(r-s)}$ is different from the $\widetilde{\mathbb{R}}^{(n-(r-s))\times(r-s)}$ before, because the constraints only contain U_1 but does not contain U_2. Fewer constraints mean a larger subspace, and we have

$$\begin{aligned}\widehat{\mathbb{R}}^{(n-(r-s))\times(r-s)} &= \left\{\Lambda_{21} \in \mathbb{R}^{(n-(r-s))\times(r-s)} : \ U_{1,2}^\top \Lambda_{21} = -U_{1,1}^\top U_{3,1}(U_{3,1} + I_{r-s})^{-1}\right\} \\ &= \widetilde{\mathbb{R}}^{(n-(r-s))\times(r-s)} + \left(\mathrm{Ker}(U_{1,2}^\top)\backslash\mathrm{Ker}(U_{X,2}^\top)\right) \\ &= \Lambda_{21}^\circ + \mathrm{Ker}(U_{X,2}^\top) + \left(\mathrm{Ker}(U_{1,2}^\top)\backslash\mathrm{Ker}(U_{X,2}^\top)\right).\end{aligned}$$

Let $h^{(2)}$ be the chart function for the extra subspace, then

$$(h, h^{(2)}): \quad \widehat{\mathbb{R}}^{(n-(r-s))\times(r-s)} \to \mathbb{R}^{(n-(r-s))(r-s)-r(r-s)} \oplus \mathbb{R}^{(r-s)(r-s)},$$
$$\Lambda_{21} \mapsto (\Gamma, \Gamma^{(2)}).$$

Putting them together, we have

$$(\mathrm{id}, h, h^{(2)}) \circ g \circ f: \quad (P_1, U_3) \mapsto (\Omega_{11}, \Omega_{21}, \Lambda_{11}, \Gamma, \Gamma^{(2)}).$$

The chart function φ is a coordinate slice of the above mapping.

It suffices to find the chart functions for the remaining components of $\widetilde{f}(U,S)$, i.e., the components M_4, $\tilde{S}$, R_1, R_2. For $\tilde{S} \in \mathbb{S}_r$ and $R_1 \in \mathrm{upper}(s,s)$, the domains are Euclidean spaces with natural bases. We now look at M_4 and R_2.

Decompose M_4 into parts that are parallel to and perpendicular to the subspace of U_3:

$$M_4 = M_4^{\parallel} + M_4^{\perp}, \quad \text{where } M_4^{\parallel} = P_{U_3} M_4, \quad M_4^{\perp} = P_{U_3}^{\perp} M_4.$$

Let $M_4^{\perp} = U_4 R_4$ be the QR decomposition of $M_4^{\perp}$. Then the whole M_4 could be written as

$$M_4 = (U_3, U_4)\begin{pmatrix} R_3 \\ R_4 \end{pmatrix}, \qquad R_3 \in \mathbb{R}^{(r-s)\times(r-s)}, \quad R_4 \in \ \mathrm{upper}(s,s).$$

In this way, we can re-parameterize (M_4, R_2) using (U_4, R_2, R_3, R_4):

$$p: (M_4, R_2 - I_{r-s}) \mapsto (U_4, R_2, R_3, R_4).$$

The domain of U_4 is the constrained Stiefel manifold $\widetilde{\mathrm{St}}(n, s; U_1, U_3)$. Just as before, we can construct a composite function for this constrained Stiefel manifold:

$$\begin{aligned}
g^{(3)} : \quad \widetilde{\mathrm{St}}(n, s; U_1, U_3) &\to \mathrm{Skew}(s) \oplus \widetilde{\mathbb{R}}^{(n-s)\times s}, \\
U_4 &\mapsto (\Pi_{11}, \Pi_{21}); \\
h^{(3)} : \quad \widetilde{\mathbb{R}}^{(n-s)\times s} &\to \mathbb{R}^{(n-s)s-rs}, \\
\Pi_{21} &\mapsto \Xi; \\
(\mathrm{id}, h^{(3)}) \circ g^{(3)} : \quad \widetilde{\mathrm{St}}(n, s; U_1, U_3) &\to \mathrm{Skew}(s) \oplus \mathbb{R}^{(n-s)s-rs}, \\
U_4 &\mapsto (\Pi_{11}, \Xi).
\end{aligned}$$

The remaining components are R_2, R_3, and R_4. The constraints for them come from the requirement that U as a whole is in $\mathrm{St}(n, r)$. This gives

$$\begin{aligned}
U^\top U &= \Big((U_1R_1, 0) + (M_4, U_3R_2)\Big)^\top \Big((U_1R_1, 0) + (M_4, U_3R_2)\Big) \\
&= \begin{pmatrix} R_1^\top U_1^\top U_1 R_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \left((U_4, U_3)\begin{pmatrix} R_4 & 0 \\ R_3 & R_2 \end{pmatrix}\right)^\top \left((U_4, U_3)\begin{pmatrix} R_4 & 0 \\ R_3 & R_2 \end{pmatrix}\right) \\
&= \begin{pmatrix} R_1^\top R_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} R_4 & 0 \\ R_3 & R_2 \end{pmatrix}^\top \begin{pmatrix} R_4 & 0 \\ R_3 & R_2 \end{pmatrix} \\
&= \begin{pmatrix} R_1^\top R_1 + R_3^\top R_3 + R_4^\top R_4 & R_3^\top R_2 \\ R_2^\top R_3 & R_2^\top R_2 \end{pmatrix} = I_r.
\end{aligned}$$

Denote

$$R_0 := \begin{pmatrix} R_2 & R_3 \\ \mathbf{0} & R_4 \end{pmatrix} \in \ \mathrm{upper}(r, r).$$

Then the $r \times r$ upper-triangular matrix R_0 should satisfy

$$R_0^\top R_0 = \begin{pmatrix} I_{r-s} & \mathbf{0} \\ \mathbf{0} & I_s - R_1^\top R_1 \end{pmatrix}.$$

Such R_0 is uniquely determined. Therefore, we get the following chart function for the components (M_4, R_2):

$$\begin{aligned}
(\mathrm{id}, h^{(3)}) \circ g^{(3)} \circ p : \quad \widetilde{\mathbb{R}}(n, s) \oplus \widetilde{\mathrm{upper}}(r-s, r-s) &\to \mathrm{Skew}(s) \oplus \mathbb{R}^{(n-s)s-rs}, \\
(M_4, R_2 - I_{r-s}) &\mapsto (\Pi_{11}, \Xi).
\end{aligned}$$

Putting everything together, we get the following chart function for the manifold $\mathcal{M}$:

$$\begin{aligned}\widetilde{\varphi} := \Big((\mathrm{id}, h, h^{(2)}) \circ g,\ (\mathrm{id}, h^{(3)}) \circ g^{(3)} \circ p,\ \mathrm{id}\Big) \circ \widetilde{f} :&\\ \mathcal{M} \to \Big(\mathrm{Skew}(s) \oplus \mathbb{R}^{(r-s)s} \oplus \mathrm{Skew}(r-s) \oplus \mathbb{R}^{(n-2r+s)(r-s)} \oplus \mathbb{R}^{(r-s)(r-s)}\Big)&\\ \oplus \Big(\mathrm{Skew}(s) \oplus \mathbb{R}^{(n-s-r)s}\Big) \oplus \mathbb{S}_r \oplus \mathrm{upper}(s,s),&\\ (U, S) \mapsto \left(\Big(\Omega_{11}, \Omega_{21}, \Lambda_{11}, \Gamma, \Gamma^{(2)}\Big), \Big(\Pi_{11}, \Xi\Big), \widetilde{S} - \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix}, R_1 - I_s\right).&\end{aligned}$$

The chart function φ is a coordinate slice of the chart function $\widetilde{\varphi}$. Hence, $\mathcal{N}_{Z\#}$ is an embedded submanifold of $\mathcal{M}$. □

Chapter 5

FAST GLOBAL CONVERGENCE FOR LOW-RANK MATRIX RECOVERY

This chapter presents the central result of this thesis, which is the global convergence guarantee for a class of low-rank matrix recovery problems under a unified framework. We show that for the population least squares loss function, under certain assumptions, with high probability the Riemannian gradient descent on the manifold $\mathcal{M}_r$ starting from a global random initialization converges to the ground truth in a nearly linear convergence rate, i.e., it takes $O(\log \frac{1}{\epsilon} + \log n)$ iterations to reach an ϵ-accurate solution.

Our result is the first to establish a nearly optimal and almost dimension-free convergence rate for general rank-r problems. In this sense, it fully explains the behavior of the Riemannian gradient descent trajectory observed in numerical experiments. This is in contrast to previous works on the global convergence guarantee with random initialization for such low-rank recovery problems, which either only obtains a geometric convergence rate that is slower than ours, or only tackles the case where the rank $r = 1$, see the comparison in Section 1.3.

By analyzing the low-rank recovery problems on the low-rank matrix manifold $\mathcal{M}_r$, it brings about unique benefits and challenges at the same time. On the one hand, the Łojasiewicz inequality can be satisfied on the manifold, and this naturally leads to linear convergence. On the other hand, the spurious critical points $\mathcal{S}_\#$ (Section 2.3) and their local neighborhoods pose significant challenges to the analysis because the spurious critical points have singular Riemannian gradient and the Łojasiewicz inequality can be violated near them.

To tackle this difficulty, we conduct a careful analysis of the spurious regions near the spurious critical points. We divide the trajectory into three stages, and show the preservation of the concentration results of the initialization throughout the trajectory. We show that with high probability, the trajectory either avoids the spurious regions, or enters and escapes from them in a small number of iterations.

Organization of this chapter. We have given a detailed introduction of the problem setting and related work in Section 1.3. The rest of this chapter is organized as follows. In Section 5.1, we present the main result of this chapter, namely the nearly linear convergence guarantee for the randomly initialized Riemannian gradient descent toward the ground truth when minimizing the population least squares loss function. The next few sections are devoted to the proof of the main result. More specifically, Section 5.2 outlines the key intermediate steps to describe the trajectory. Section 5.3 introduces the Łojasiewicz convergence tool as a fundamental convergence guarantee. Section 5.4 discusses the spurious regions around the spurious critical points. Section 5.5 gives some technical lemmas about the dynamics of the trajectory behavior. Section 5.6 presents the proofs for the main result in Section 5.1. The next few sections discuss a different population loss function that stems from the phase retrieval problem. Section 5.7 gives the main results for this loss function, and Section 5.8 presents the proofs. Finally, in Section 5.9, we make some concluding remarks.

5.1 Main results

In this section, we introduce the main results of this chapter. The optimization problems we are looking at are (1.2) and (1.3), namely we minimize either $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$ or $F_2(Z) = \frac{1}{2}(\|Z\|_F - \|X\|_F)^2 + c\|Z - X\|_F^2$ on the low-rank matrix manifold $\mathcal{M}_r$. The analysis focuses on the case of symmetric positive semi-definite (SPSD) or Hermitian positive semi-definite (HPSD) matrices, and one may refer to [65] for potential extension to the asymmetric case.

Preliminaries

We first introduce the necessary notations, optimization techniques, and assumptions for the statement of our main results.

Notations. Throughout the chapter, $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, and we use $(\cdot)^*$ to denote the real transpose when $\mathbb{F} = \mathbb{R}$ and the Hermitian transpose when $\mathbb{F} = \mathbb{C}$. Denote $\mathcal{M}_r$ as the fixed rank manifold $\{Z \in \mathbb{F}^{n_1 \times n_2} : \operatorname{rank}(Z) = r\}$. In the SPSD/HPSD case, $\mathcal{M}_r = \{Z \in \mathbb{S}_n \text{ or } \mathbb{H}_n : \operatorname{rank}(Z) = r\}$, where $\mathbb{S}_n$ or $\mathbb{H}_n$ denotes the set of n by n real symmetric or Hermitian matrices. Let $\overline{\mathcal{M}_r}$ denote the closure of $\mathcal{M}_r$. We always use X to denote the ground truth matrix of size $n \times n$ and rank r, while we use Z_* to denote any fixed point or limit point. Let $X = UDU^*$ be the eigenvalue decomposition of the ground truth matrix X, where $D = \text{diag}\{d_1, \ldots, d_r\}$ is in descending order unless otherwise specified. Let $Z = V_z \Sigma_z V_z^*$ be the eigenvalue decomposition of a matrix Z, where $\Sigma_z = \text{diag}\{\sigma_1, \ldots, \sigma_r\}$ is also in descending order unless otherwise specified. In general, we use $\sigma_j(\cdot)$ to denote the j-th largest eigenvalue of a matrix. Let $\kappa := \frac{d_1}{d_r}$ denote the condition number of the ground truth X. For an integer $s > 0$, let $[s] = \{1, 2, ..., s\}$. For any $j \in [r]$, let $\backslash j := [r] \setminus \{j\}$. The vector norm we use is $\| \cdot \|_2$ and the matrix norm is $\| \cdot \|_F$ unless otherwise specified. We use $0 < c, C < \infty$ to denote any absolute constant independent of n in our statement. These constants may vary in different contexts. The symbol $\Omega(n)$ means that there exist constants $C \geq c > 0$ such that $c \cdot N \leq \Omega(n) \leq C \cdot n$, and $O(n)$ means that there exist a constant $C > 0$ such that $O(n) \leq C \cdot n$. In this chapter, we focus on the large n regime, and other quantities including r, σ_r and d_r will be treated as constants and will be ignored in the $O(\cdot)$ and $\Omega(\cdot)$ notations. The symbol poly(a) stands for a nonnegative quantity upper bounded by $C \cdot a^k$ for some $C > 0$ and $k \in \mathbb{N}_+$.

Assumptions 5.1.1. *The following assumptions are necessary for the main results:*

Assumption 1: Assume $\alpha > 0$ is a small constant such that the discretized system can be well approximated by the continuous system. In other words, let $A(t)$ denote a continuous system of interest and $\{A_k\}$ be its time discretization, then we assume that $A_{k+1} = A_k + \alpha \dot{A}_k + o(\alpha \dot{A}_k)$.

Assumption 2: Assume that the eigenvalues of Z are always simple and do not cross one another along the whole gradient flow or gradient descent trajectory. Moreover, the smallest eigenvalue gap is lower bounded along the whole trajectory, i.e., $|\sigma_i(Z) - \sigma_j(Z)| \geq c_g \max\{\sigma_i(Z), \sigma_j(Z)\}$ for any $i, j \in [r]$ and $i \neq j$.

Remark 5.1.2. The gradient flow of Z with non-crossing singular values is common in practice. In fact, it is *generic*[1] as is proved in [49]. However, gradient flows with crossing singular values have also been observed in some experiments. Assumption 2 is mainly for the purpose of simplifying the presentation of our technical analysis, as crossing singular values would introduce additional difficulties. These additional difficulties can be poten-

[1] A property of a topological space is called *generic* if it holds for a subset of the space which is of the second Baire category.

tially overcome by considering subspaces of singular vectors as a whole. This is left for future work.

Statement of main result

We will prove that the randomly initialized Riemannian gradient descent for $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$ with high probability escapes the set of spurious critical points $\mathcal{S}_\#$ (defined in (5.1)) and converges to the global minimum X. With high probability, the convergence rate is nearly linear and almost dimension-free and condition number free. By choosing a small enough constant step size, for any $\epsilon > 0$, with high probability it only takes $O(\log n + \log \frac{1}{\epsilon})$ iterations to get an ϵ-accurate solution. More specifically, we have the following theorem.

Theorem 5.1.3. *Consider $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$, and let $\{Z_k\}_{k=0}^{\infty}$ be the sequence generated by the Riemannian gradient descent initialized at Z_0 which is drawn from the general random distribution (as defined in Definition 5.5.1). Under Assumptions 5.1.1, there exists constant stepsize α, such that with high probability no less than $1 - \frac{1}{poly(\log n)}$, we have $\lim_{k\to\infty} Z_k = X$. Furthermore, there exists an integer $\mathcal{K} = O(\frac{1}{\alpha}\log n)$ such that:*

$$\|Z_k - X\|_F^2 \leq e^{-ck}, \textit{ for all } k \geq \mathcal{K}.$$

Here $c = \Omega(\alpha) > 0$. In other words, it takes no more than $O(\frac{1}{\alpha}\cdot(\log\frac{1}{\epsilon} + \log n))$ iterations to get an ϵ-accurate solution (i.e., $\|Z - X\| \leq \epsilon\|X\|_F$) via the randomly initialized Riemannian gradient descent.

The proof of Theorem 5.1.3 consists of three stages, described by Theorem 5.2.1, Theorem 5.2.2, and Theorem 5.2.3 respectively. We will provide more details of the proof strategy in the next few sections. Below we give some high-level explanation of what happens in those three stages.

Sketch of proof

Here we highlight a few high-level ideas of our proofs for the main theorem stated in the previous subsection. A more comprehensive and detailed presentation of the proof strategy and intermediate results can be found in the next few sections.

Fundamental convergence guarantee by the Łojasiewicz inequality. The Łojasiewicz inequality [3, 9, 21, 118] has long been studied as a fundamental tool for convergence analysis. It is especially useful in proving the linear convergence rate of first-order optimization methods. In Section 5.3, we derive a version of this tool tailored for the Riemannian gradient descent method, stated in Theorem 5.3.1. Using this theorem, the task of checking the convergence rate for a specific problem is reduced to checking conditions (D) and (L) for the objective function. We then observe that for problems (1.2) and (1.3), these conditions are satisfied except for some small regions on the manifold. Such regions are later dubbed *spurious regions*, illustrated by the blue balls in Figure 5.1. The spurious regions lead to the next important result on the geometry of the low-rank matrix manifold.

Geometry of spurious regions over $\mathcal{M}_r$. We have discussed in Section 2.3 that for the simple least squares loss function $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$, apart from the ground truth solution $Z = X$, there are some *spurious critical points* in $\overline{\mathcal{M}_r} \backslash \mathcal{M}_r$, which we denote as $\mathcal{S}_\#$:

$$\mathcal{S}_\# := \left\{Z_* : \; Z_* = U_1 D_1 U_1^*, \; \text{where } U = (U_1, U_2)\,, \; D = \operatorname{diag}\,\{D_1, D_2\}, \; Z_* \neq X\right\}. \tag{5.1}$$

See Lemma 2.3.1 for more details about the derivation of spurious critical points. If X has distinct eigenvalues, then $|\mathcal{S}_\#| = 2^r - 1$. In particular, every point in $\mathcal{S}_\#$ has local geometry somewhat similar to that of a saddle point, except that the Riemannian gradient is singular. Formally, the Riemannian Hessian has a $-\infty$ direction, which indicates that the curvature is singular at such point.

In the local neighborhoods of these spurious critical points, there are certain *spurious regions* where Condition (L) could fail. In other words, a linear convergence rate is guaranteed outside the spurious regions, while the convergence rate slows down inside the spurious regions. It then becomes important to characterize how the Riemannian gradient descent may escape these regions and converge to the ground truth. A full description of the spurious regions, along with examples and illustrations, is given in Section 5.4.

Three-stage description of the trajectory behavior. It now remains to study how the trajectory of the randomly initialized gradient descent sequence escapes the spurious regions and converges to the ground truth on the manifold. In Section 5.2, we divide the whole trajectory into three stages. In the worst case, the sequence is dragged toward $Z_* = 0$ in the first stage, then escapes $Z_* = 0$ and approaches some other $Z_* \in \mathcal{S}_\# \setminus \{0\}$ in the second stage, and finally escapes such Z_* in the third stage. We show that by throwing out a small probability measure, the total number of iterations needed to reach the ϵ-neighborhood of X is bounded by $O(\log \frac{1}{\epsilon} + \log n)$, as stated in the main theorems. This worst case is, however, rarely observed in numerical experiments; usually, the sequence generated by the Riemannian gradient descent simply avoids all spurious critical points quickly and converges to the ground truth. Proofs of the three-stage behavior are stated in Theorem 5.2.1, Theorem 5.2.2, Theorem 5.2.3, and they further depend on a series of technical lemmas detailed in Section 5.5.

More specifically, from Lemma 5.5.3, the "angle" (the product of the two column vector matrices) between the randomly initialized column space and the ground truth column space is of order $\Omega(\arccos \frac{1}{\sqrt{n}})$. If the angle remains less than $\Omega(\arccos \frac{1}{\text{poly}(n)})$ when the sequence enters the spurious region $\mathcal{B}(\mathcal{S}_\#, \delta)$ (defined in Lemma 5.4.1) and it grows exponentially fast, it only takes $O(\log n)$ iterations for the angle to become $\Omega(1)$, which means the sequence has escaped the spurious region. Note that the spurious critical point Z_* differs from a classical strict saddle in that the Hessian of Z_* has $-\infty$ directions. Still, we can bound the number of iterations needed to escape from $\mathcal{B}(\mathcal{S}_\#, \delta)$ by throwing out a small probability measure. Details can be found in the proof of Theorem 5.2.3.

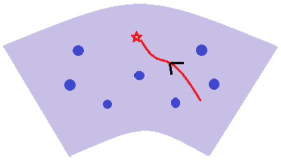


Figure 5.1: Illustration of the trajectory of the Riemannian gradient descent on $\mathcal{M}_r$.

5.2 Key intermediate steps for Theorem 5.1.3

This section and the upcoming sections are primarily devoted to the proof of Theorem 5.1.3. This theorem is the central result of this chapter and fully describes the trajectory behavior of the Riemannian gradient descent for the least squares loss function on the rank-r matrix manifold $\mathcal{M}_r$.

To prove Theorem 5.1.3, we divide the trajectory of the Riemannian gradient descent on the manifold into three stages.

Stage 1: The iterative sequence generated by the Riemannian gradient descent starts from a random initialization point. Then, with high probability, the sequence will enter the spurious region of $Z_* = 0$, where the radius of the spurious region δ is a constant satisfying $\delta = \Omega(1)$.

Stage 2: Following stage 1, as the sequence reaches the δ-spurious region of $Z_* = 0$, it further takes $O(\log n)$ iterations to escape this spurious region, and enters stage 3. The sequence then stays away from the spurious region $\mathcal{B}(0, \delta)$.

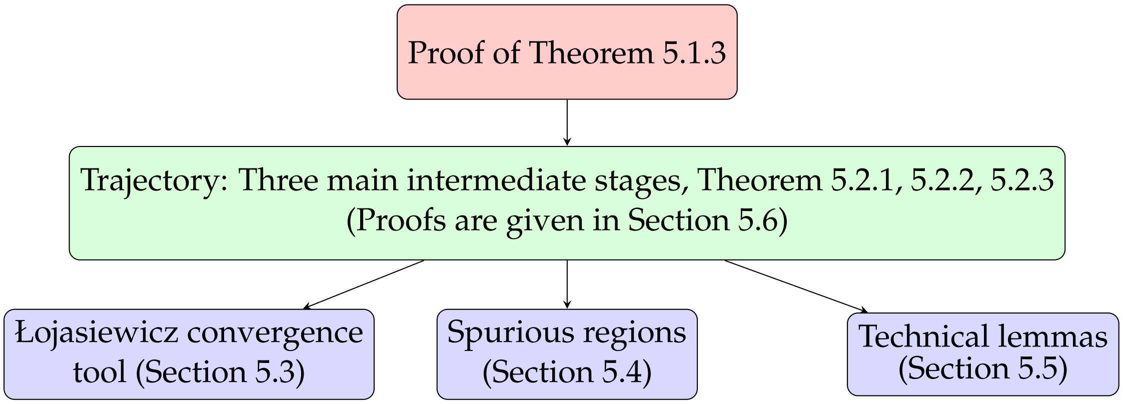
Stage 3: The sequence either enters the δ-local region of the ground truth X without entering any spurious region of $Z_* \in \mathcal{S}_\# \setminus \{0\}$, or reaches a δ-spurious region of some $Z_* \in \mathcal{S}_\# \setminus \{0\}$. If it enters the δ-local region of the ground truth X, then it takes $O(\log \frac{1}{\epsilon})$ iterations to reach an ϵ-accurate solution. On the other hand, if it reaches a δ-spurious region of some $Z_* \in \mathcal{S}_\# \setminus \{0\}$, then with high probability, after $O(\text{poly}(\log n))$ iterations the sequence will escape the δ-spurious region.

The following three theorems state the above three stages in mathematical terms respectively. Their proofs are given in Section 5.6.

Theorem 5.2.1. *Under the setting of Theorem 5.1.3, for the sequence initialized at Z_0 and generated by the Riemannian gradient descent, there exists $K_0 = \Omega(1)$ such that with high probability exceeding $1-\frac{1}{poly(n)}$, we have $\cup_{k\le K_0} Z_k \notin \cup_{Z_* \in \mathcal{S}_\# \setminus \{0\}} \mathcal{B}(Z_*, \delta)$ and $Z_{K_0} \in \mathcal{B}(0, \delta)$, i.e., the sequence enters $\mathcal{B}(0, \delta)$ at step K_0. Here, $\delta = \Omega(1)$ is a constant that depends only on X, the constants C_1 and C_2 in the general random distribution (Definition 5.5.1), and the step size α, and $K_0 = O(\log \delta) = \Omega(1)$.*

Theorem 5.2.2. *Under the setting of Theorem 5.1.3 and following the first stage in Theorem 5.2.1, suppose $K_0 = \Omega(1)$ is the positive integer such that $(\cup_{k\le K_0} Z_k) \cap \mathcal{B}(0, \delta) = \emptyset$ and $Z_{K_0} \in \mathcal{B}(0, \delta)$, i.e., the sequence enters $\mathcal{B}(0, \delta)$ at step K_0. Then there exists an absolute constant $C_1 = O(1) > 0$ such that with high probability exceeding $1 - \frac{1}{poly(n)}$, we have $\cup_{k\ge K_1}^{\infty} Z_k \cap \mathcal{B}(0, \delta) = \emptyset$. Here, $K_1 = K_0 + O(\log n)$.*

Theorem 5.2.3. *Under the setting of Theorem 5.1.3 and following the first and second stages in Theorem 5.2.1 and Theorem 5.2.2, assume there exists a positive* *integer K_1 and some $Z_* \in \mathcal{S}_\# \backslash \{0\}$ such that $Z_{K_1} \in \mathcal{B}(Z_*, \delta)$ and $Z_{k \le K_1} \cap \mathcal{B}(Z_*, \delta) = \emptyset$, i.e., the sequence enters the local neighborhood $B(Z_*, \delta)$ of a nonzero spurious critical point Z_* at step K_1. Then, with high probability exceeding $1 - \frac{1}{poly(\log n)}$, we have $\|Z_k - X\|_F \le \epsilon \|X\|_F$ for $k \ge K_2$. Here, $K_2 = K_1 + O(\frac{1}{\alpha} \cdot (\log n + \log \frac{1}{\epsilon}))$ and $c_* > 0$ is a constant.*


Proof of Theorem 5.1.3
Trajectory: Three main intermediate stages, Theorem 5.2.1, 5.2.2, 5.2.3
(Proofs are given in Section 5.6)
Łojasiewicz convergence tool (Section 5.3)
Spurious regions (Section 5.4)
Technical lemmas (Section 5.5)

Proof of Theorem 5.1.3. Let the stepsize α and δ be small constants that meet the requirements in Theorem 5.2.1, Theorem 5.2.2, and Theorem 5.2.3. Theorem 5.1.3 can be proved by combining the result of Theorem 5.2.1, Theorem 5.2.2 and Theorem 5.2.3. Note that if the sequence does not enter $\cup_{Z* \in \mathcal{S}_\#} \mathcal{B}(Z_*, \delta)$ in stage 3, then $\|P_{T_{Z_k}}(\nabla f(Z_k))\| \ge \delta$ always holds true after the end of stage 2, and by Lemma 5.4.5, we have that $\{Z_k\}_{k=0}$ converges to X directly in a linear rate. Otherwise, if the sequence does enter some spurious region in stage 3, we can bound the number of steps needed to converge to X by Theorem 5.2.3. □

The proofs of Theorems 5.2.1-5.2.3 follow the roadmap laid out in Section 5.1. In particular, from Section 5.3 to 5.5, each subsection introduces a group of technical results corresponding to a main idea in Section 5.1. They are the fundamental convergence tool by the Łojasiewicz inequality, the geometry of spurious regions on $\mathcal{M}_r$, and the technical lemmas describing the dynamics of the trajectory behavior. We then use those technical results to prove Theorems 5.2.1–5.2.3 in Section 5.6.

5.3 Fundamental convergence guarantee of the Riemannian gradient descent

The *Łojasiewicz inequality*, which is named after S. Łojasiewicz [100, 101], is a powerful tool for analyzing the convergence rate of gradient-based methods. Previous works have used the Łojasiewicz inequality to prove the convergence rate in many Euclidean optimization problems as well as Riemannian optimization problems, see for example [3, 8–10, 21, 118, 128].

The following theorem serves as a primary tool to determine the convergence rate of the Riemannian gradient descent when minimizing a differentiable function $f : \mathcal{M} \to \mathbb{R}$. We assume that $\{Z_k\}$ generated by the Riemannian gradient descent is bounded for the rest of the chapter.

Theorem 5.3.1. *Let $\mathcal{M}$ be a Riemannian manifold, $f : \mathcal{M} \to \mathbb{R}$ be a differentiable loss function to be minimized, $\{Z_k\}_{k=0}^{\infty}$ be a sequence generated by the Riemannian gradient descent algorithm* (2.6)*. Assume that the following conditions hold:*

1) (Descent Inequality) *There exists $C_d > 0$ such that*

$$f(Z_k) - f(Z_{k+1}) \geq C_d \|P_{T_{Z_k}}(\nabla f(Z_k))\| \|Z_{k+1} - Z_k\|; \tag{D}$$

2) (Łojasiewicz Gradient Inequality) *There exists $0 < C_l < +\infty$ such that*

$$|f(Z_k) - f(Z_*)|^{1-\omega} \leq C_l \|P_{T_{Z_k}}(\nabla f(Z_k))\|, \tag{L}$$

with $0 < \omega \leq \frac{1}{2}$. Here, Z_* is the accumulating point of $\{Z_k\}$.

Then, if the learning rate α satisfies $0 < \alpha < \frac{2C_l^2}{C_d}$, the sequence $\{Z_k\}$ converges to its accumulating point Z_* with the following convergence rate:

$$\|Z_k - Z_*\| \leq \begin{cases} e^{-ck} & \text{, if } \omega = \frac{1}{2}; \\ k^{-\frac{\omega}{1-2\omega}} & \text{, if } 0 < \omega < \frac{1}{2}. \end{cases}$$

When $\omega = \frac{1}{2}$, linear convergence rate can be guaranteed with $c = -\log(1 - \frac{\alpha C_d}{2C_l^2}) > 0$. Here $\|\cdot\|$ is an arbitrary norm under which there is a first-order retraction on the manifold.

Proof. Since $\{f_k\}_{k=0}^{\infty}$ is a monotone and lower bounded sequence, Condition (D) and continuity of $f(\cdot)$ implies convergence of $\{Z_k\}$ to some fixed point Z_*. Without loss of generality we assume that $f(Z_*) = 0$. By Conditions (D) and (L), we have

$$\begin{aligned} \|Z_{k+1} - Z_k\|_F &\leq \frac{1}{C_d \|P_{T_z}(\nabla f(Z_k))\|_F}(f_k - f_{k+1}) \leq \frac{C_l}{C_d} f_k^{\omega-1}(f_k - f_{k+1}) \\ &\leq \frac{C_l}{C_d} \int_{f_{k+1}}^{f_k} \phi^{\omega-1} d\phi = \frac{C_l}{\omega C_d}(f_k^{\omega} - f_{k+1}^{\omega}). \end{aligned}$$

Since $\{f_k\}_{k=0}^{\infty}$ is a monotone and lower bounded sequence, $\{f_k\}$ is convergent. Therefore, $\{Z_k\}$ is convergent, and the limit point is Z_*.

Consider $s_k := \sum_{i=k}^{\infty} \|Z_{i+1} - Z_i\|_F \leq \frac{C_l}{\omega C_d} f_k^{\omega}$. Combined with Condition (L), we get

$$s_k^{\frac{1-\omega}{\omega}} \leq (\frac{C_l}{\omega C_d})^{\frac{1-\omega}{\omega}} f_k^{1-\omega} \leq C_l (\frac{C_l}{\omega C_d})^{\frac{1-\omega}{\omega}} \|P_{T_{Z_k}}(\nabla f(Z_k))\|_F.$$

Let $\xi_k = -P_{T_{Z_k}}(\nabla f(Z_k))$ and $\alpha\widetilde{\xi_k} = Z_{k+1} - Z_k = \mathcal{R}(Z_k + \alpha\xi_k) - Z_k$. By the first-order retraction property, $\alpha\widetilde{\xi_k} = \alpha\xi_k + o(\alpha\|\xi\|)$. This gives us

$$s_k^{\frac{1-\omega}{\omega}} \leq C_l (\frac{C_l}{\omega C_d})^{\frac{1-\omega}{\omega}} \|\xi_k\|_F = C_l (\frac{C_l}{\omega C_d})^{\frac{1-\omega}{\omega}} \frac{1}{\alpha} (\|Z_{k+1} - Z_k\|_F + o(\|Z_{k+1} - Z_k\|_F)).$$

Let $\rho = \rho(\alpha) := \frac{\alpha}{C_l} (\frac{\omega C_d}{C_l})^{\frac{1-\omega}{\omega}}$, i.e., ρ is a constant depending only on C_l, C_d and ω. Then we have

$$\rho(\alpha) s_k^{\frac{1-\omega}{\omega}} = \|Z_{k+1} - Z_k\|_F + o(\|Z_{k+1} - Z_k\|_F).$$

This implies

$$\rho(\alpha) s_k^{\frac{1-\omega}{\omega}} \leq (1 + o(1))(s_k - s_{k+1}).$$

By choosing α to be a small enough constant, there exists $\rho \in (0, 1)$ such that

$$\rho s_k^{\frac{1-\omega}{\omega}} \leq s_k - s_{k+1}. \tag{5.2}$$

In the case of $\omega = \frac{1}{2}$, the above inequality gives $s_{k+1} \leq (1 - \rho)s_k$ for some $\rho \in (0, 1)$. This implies $\|Z_k - Z_*\|_F \leq s_k \leq (1 - \rho)^k s_0 = s_0 e^{-ck}$, where $c = -\log(1 - \rho) = -\log(1 - \frac{\alpha C_d}{2C_l^2}) > 0$, which implies linear convergence rate.

On the other hand, in the case of $0 < \omega < \frac{1}{2}$, assuming there exists p such that $s_k = c_1 k^{-p}$, we solve for p as follows:

$$\begin{aligned} s_{k+1} &= c_1 \frac{1}{(k+1)^p} = c_1 \frac{1}{k^p}(1 + \frac{1}{k})^{-p} \\ &= c_1 \frac{1}{k^p}(1 - \frac{p}{k}) + O(\frac{1}{k^{p+2}}) \\ &= s_k(1 - p s_k^{\frac{1}{p}} c_1^{-\frac{1}{p}}) + O(\frac{1}{k^{p+2}}) \\ &\leq s_k(1 - \rho s_k^{\frac{1-2\omega}{\omega}}), \end{aligned}$$

where the last inequality follows from Equation (5.2). Choose c_1 a proper constant, then the above inequality holds with $p = \frac{\omega}{1-2\omega}$. Thus, we obtain $\|Z_k - Z_*\|_F \leq s_k \leq c_1 k^{-\frac{\omega}{1-2\omega}}$. This implies polynomial convergence rate. □

Remark 5.3.2. A few remarks are in order.

1) This theorem explores the convergence rate of the Riemannian gradient descent to an accumulating point. This convergence rate depends on the property of the function f (reflected in the exponent ω), the constants C_d and C_l, and the learning rate $\alpha > 0$.

2) With Theorem 5.3.1, the task of checking the convergence rate is reduced to checking conditions (D) and (L) and determining the respective constants.

3) This theorem only requires Z_* to be an accumulating point, but Z_* is not necessarily a local minimum. It can also be other types of critical points such as saddle points.

4) From this result, we can see that the convergence rate is faster with a larger stepsize α, or a larger Riemannian gradient $\|P_{T_{Z_k}}(\nabla f(Z_k))\|_F$ (which makes C_l smaller).

5) An extension of this theorem has been proposed in [134], which allows for mixed norms in (D) and (L), see also Chapter 7.

Lemma 5.3.3. *Let $f(Z) = F_1(Z)$ or $F_2(Z)$. Assume that Z_k satisfies Conditions* (D) *and* (L) *in Theorem 5.3.1* only *for $k = 1, \dots, K$, with $K < +\infty$. Then, we have*

$$f_{k+1} < (1-\rho) f_k, \quad k = 1, \dots, K,$$

where $f_k := f(Z_k)$. Furthermore, if $\alpha > 0$ is small enough, we have $\|Z_k - X\|_F \lesssim e^{-ck}$, where $\rho = \Omega(\alpha \frac{C_d}{C_l^2}) \in (0,1)$ and $c = -\frac{1}{2}\log(1-\rho) > 0$.

Proof. For any finite $k \in [K]$, from Conditions (D) and (L), by the first-order retraction property, we have:

$$\begin{aligned} f_{k+1} &\leq f_k - C_d \|P_{T_{Z_k}}(\nabla f(Z_k))\| \|Z_{k+1} - Z_k\| \\ &\leq f_k - \Omega(\alpha) C_d \|P_{T_{Z_k}}(\nabla f(Z_k))\|^2 \\ &\leq \left(1 - \Omega(\alpha)\frac{C_d}{C_l^2}\right) f_k. \end{aligned}$$

Since for both $F_1(Z)$ and $F_2(Z)$, we have $\|Z_k - X\|_F^2 \lesssim f_k$, we have $\|Z_k - X\|_F \lesssim e^{-ck}$, with $\rho = \Omega(\alpha \frac{C_d}{C_l^2})$ and $c = -\frac{1}{2}\log(1-\rho)$. If $\alpha > 0$ is properly small, we have $c > 0$. □

The two conditions stated in Theorem 5.3.1 are satisfied in a large part of the manifold $\mathcal{M}_r$ for the loss functions that we consider, including the neighborhood of the ground truth X. The fast convergence rate in most parts of the manifold is thus easy to derive. In some regions of the manifold, though, Condition (L) could fail and the convergence rate could deteriorate. As we will see in the next section, these regions are the *spurious regions* near the *spurious critical points* on the manifold. Special analysis is needed to study how the Riemannian gradient descent escapes these spurious regions.

We also mention that another version of Theorem 5.3.1 is given in Section 5.8, specially tailored for the objective functions with weak isometry, which will be used in our convergence study of the Riemannian gradient descent for phase retrieval.

5.4 Geometry of the spurious regions on the low-rank matrix manifold

The spurious critical points play an important role in the analysis of the convergence guarantee for $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$. In addition to the spurious critical points themselves, here we need to look at some local regions in their δ-neighborhoods, which we call the *spurious regions* and denote as $\mathcal{B}(\mathcal{S}_\#, \delta)$. This is where Condition (L) of Theorem 5.3.1 is violated. In the next two sections, we show that with high probability, the trajectory either avoids the spurious regions, or escapes from them in a small number of steps.

Spurious regions in the neighborhood of spurious fixed points

As is mentioned in Section 5.3, there are some regions on $\mathcal{M}_r$ that violate the conditions for fast convergence guarantee in Theorem 5.3.1. These are the regions near the spurious critical points. We have discussed the motivation and definition of the spurious critical points in Section 2.3. In this subsection, we take a detailed look at the spurious regions near those spurious critical points.

To ensure that Condition (L) holds with $\omega = \frac{1}{2}$, what we essentially need is

$$\|P_{T_z}(Z - X)\|_F \geq C_L, \qquad C_L > 0.$$

However, for some special Z which occupies a small part of the whole domain, the Riemannian gradient $P_{T_z}(Z - X)$ becomes so small that this lower bound is violated. We use *spurious regions* to refer to the regions where Z violates this lower bound.

The following results show that the spurious regions are in the neighborhoods of the spurious fixed points.

Lemma 5.4.1 (Spurious regions)**.** *The spurious regions on $\mathcal{M}_r$ can be characterized as follows:*

$$\mathcal{B}(\mathcal{S}_\#, \delta) := \cup_{Z_* \in \mathcal{S}_\#} \mathcal{B}(Z_*, \delta) = \{Z : \|P_{T_z}(Z - X)\|_F \le \delta\}.$$

More specifically, each $\mathcal{B}(Z_*, \delta)$ can be characterized as follows. In the general non-Hermitian case, assume $X = UDV^*$ is the SVD of X, then

$$\mathcal{B}(Z_*, \delta) = \left\{ Z : \, Z = \left(B, \widetilde{B}\right) \begin{pmatrix} D_1 + \Delta_1 & 0 \\ 0 & \Delta_2 \end{pmatrix} \begin{pmatrix} C^* \\ \widetilde{C}^* \end{pmatrix}, \right.$$
$$\left. \|\Delta_1\|, \, \|\Delta_2\|, \, \|P_B - P_{U_1}\|, \, \|P_{\widetilde{B}} - P_{\widetilde{U}}\|, \, \|P_C - P_{V_1}\|, \, \|P_{\widetilde{C}} - P_{\widetilde{V}}\| \le \mathcal{O}(\delta) \right\},$$

where $P_B = BB^*$ is the projection onto the subspace of B (similar for P_{U_1}, $P_{\widetilde{B}}$, $P_{\widetilde{U}}$, and those of C and V); $U = (U_1, U_2)$ is an $(s, r - s)$ dimensional splitting for some $0 < s < r$ (similar for V); $\widetilde{U}$ satisfies $\tilde{U} \subset col(U)^\perp$ and $\widetilde{U}^* \widetilde{U} = I_{r-s}$, $D = diag(D_1, D_2)$ where D_1 and D_2 are diagonal matrices, and Δ_1, Δ_2 are also diagonal matrices.

In the SPSD/HPSD case, assume that $X = UDU^*$ is the eigenvalue decomposition of X, then

$$\mathcal{B}(Z_*, \delta) = \left\{ Z : \, Z = \left(B, \widetilde{B}\right) \begin{pmatrix} D_1 + \Delta_1 & 0 \\ 0 & \Delta_2 \end{pmatrix} \begin{pmatrix} B^* \\ \widetilde{B}^* \end{pmatrix}, \right.$$
$$\left. \|\Delta_1\|, \, \|\Delta_2\|, \, \|P_B - P_{U_1}\|, \|P_{\widetilde{B}} - P_{\widetilde{U}}\| \le \mathcal{O}(\delta) \right\}.$$

Proof. We only prove it for the general non-Hermitian case. The proof of the SPSD/HPSD case is very similar and we omit the details here.

Assume $Z = U_z \Sigma_z V_z^*$, and let $\widetilde{U}_z, \widetilde{V}_z \in \mathbb{F}^{n \times (n-r)}$ be the orthogonal complements of U_z and V_z. We can express X in the following block form under this new basis:

$$X = \begin{pmatrix} U_z & \widetilde{U}_z \end{pmatrix} \widetilde{X} \begin{pmatrix} V_z^* \\ \widetilde{V}_z^* \end{pmatrix} = \begin{pmatrix} U_z & \widetilde{U}_z \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \begin{pmatrix} V_z^* \\ \widetilde{V}_z^* \end{pmatrix},$$

where

$$\widetilde{X} = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} = Q_L \cdot D \cdot Q_R^*.$$

Then we have

$$P_{T_z}(X) = \begin{pmatrix} U_z & \widetilde{U}_z \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & 0 \end{pmatrix} \begin{pmatrix} V_z^* \\ \widetilde{V}_z^* \end{pmatrix}, \quad U = \begin{pmatrix} U_z & \widetilde{U}_z \end{pmatrix} Q_L, \quad V = \begin{pmatrix} V_z & \widetilde{V}_z \end{pmatrix} Q_R.$$

Assume that $\|P_{T_z}(Z - X)\|_F \leq \delta$, then

$$\|P_{T_z}(Z - X)\|_F = \left\| \begin{pmatrix} X_{11} - \Sigma_z & X_{12} \\ X_{21} & 0 \end{pmatrix} \right\|_F = \left\| \widetilde{X} - \begin{pmatrix} \Sigma_z & 0 \\ 0 & X_{22} \end{pmatrix} \right\|_F \leq \delta.$$

Let

$$\widetilde{S} := \begin{pmatrix} \Sigma_z & 0 \\ 0 & X_{22} \end{pmatrix} = \begin{pmatrix} I_r & 0 \\ 0 & P_L \end{pmatrix} \begin{pmatrix} \Sigma_z & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_r & 0 \\ 0 & P_R^* \end{pmatrix},$$

where the second equality gives the eigenvalue decomposition of the matrix $\widetilde{S}$. Then $\|\widetilde{X} - \widetilde{Z}\|_F \leq \delta$.

Using Lemma 2.4.2, we have that the eigenvalues of $\widetilde{S}$ are δ-perturbations of those of $\widetilde{X}$. Note that the eigenvalues of $\widetilde{X}$ are the same as those of X, which are $\{d_1, \ldots, d_r\} \cup \{0\}$. On the other hand, the eigenvalues of $\widetilde{S}$ are $\{\sigma_1, \ldots, \sigma_r\} \cup \{\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_{n-r}\}$, where $\{\sigma_i\}$ and $\{\widetilde{\sigma}_i\}$ are the diagonal entries of Σ_z and Σ_{22} respectively. Thus, for each σ_i, $i \in [r]$, either $|\sigma_i - d_j| = O(\delta)$ for some $j \in [r]$, or $\sigma_i = O(\delta)$. In other words, each σ_i either captures a eigenvalue of X, or is close to zero.

Now let $\mathcal{I}$ denote the set of indices of $\{d_i\}$ captured by $\{\sigma_i\}$, and $\mathcal{I}^c = [r] \backslash \mathcal{I}$. Without loss of generality, assume $|\sigma_i - d_i| = O(\delta)$ for $i \in \mathcal{I}$, i.e., their indices also match. Let $U_1 = U(:, \mathcal{I})$, $U_2 = U(:, \mathcal{I}^c)$, $V_1 = V(:, \mathcal{I})$, $V_2 = V(:, \mathcal{I}^c)$, and $D_1 = D(\mathcal{I}, \mathcal{I})$, $D_2 = D(\mathcal{I}^c, \mathcal{I}^c)$. Then $\|\Sigma_z(\mathcal{I}, \mathcal{I}) - D_1\| = O(\delta)$, and $\|\Sigma_z(\mathcal{I}^c, \mathcal{I}^c)\| = O(\delta)$.

By Lemma 2.4.1, the singular subspaces of $\widetilde{X}$ and $\widetilde{S}$ corresponding to indices $\mathcal{I}$ are δ-close. In mathematical terms, we have

$$\|P_{I(:,\mathcal{I})} - P_{Q_L(:,\mathcal{I})}\| = O(\delta), \qquad \|P_{I(:,\mathcal{I})} - P_{Q_R(:,\mathcal{I})}\| = O(\delta),$$

where I denotes the identity matrix and P denotes the projection onto the space spanned by the column vectors. Using the relations $U = (U_z, \widetilde{U}_z)Q_L$, and $V = (V_z, \widetilde{V}_z)Q_R$, we have

$$\|P_{U_z(:,\mathcal{I})} - P_{U_1}\| = O(\delta), \qquad \|P_{V_z(:,\mathcal{I})} - P_{V_1}\| = O(\delta).$$

Let $B = U_z(:, \mathcal{I})$ and $C = V_z(:, \mathcal{I})$ and we have the results regarding B and C in the lemma.

Similarly, the singular subspaces of Z corresponding to $\mathcal{I}^c$ are δ-close to some singular subspaces perpendicular to U and V. Denote them as $\widetilde{U}$ and $\widetilde{V}$ respectively. Then we obtain

$$\|P_{U_z(:,\mathcal{I}^c)} - P_{U_1}\| = O(\delta), \qquad \|P_{V_z(:,\mathcal{I}^c)} - P_{V_1}\| = O(\delta).$$

Let $\widetilde{B} = U_z(:, \mathcal{I}^c)$ and $\widetilde{C} = V_z(:, \mathcal{I}^c)$ and we have the full result. Note that by Lemma 2.4.1, the constant in the $O(\cdot)$ notation only depends on the gap between the two groups of eigenvalues, which in our case is determined by the smallest eigenvalue of X. □

Remark 5.4.2. The intuition behind Lemma 5.4.1 is that an s-dimensional principal part of Z is "almost aligned" with an s-dimensional principal part of X, and their eigenvalues are close to each other; while the other $(r-s)$-dimensional part of Z is "almost perpendicular" to the other $(r-s)$-dimensional part of X, and the eigenvalues of that part of Z is very small.

Lemma 5.4.3. *Let $X = UDU^*$ and $Z = V_z\Sigma_z V_z^*$ be the eigenvalue decompositions of X and Z respectively (here the diagonals of Σ_z are not required to be in descending order). Assume that $U = V_zR + \widetilde{V}_zS$, with $\widetilde{V}_z \in Col(V_z)^\perp$. If $Z \in \mathcal{B}(\mathcal{S}_\#, \delta)$, where $rank(Z_*) = s$, then we have*

$$R = \begin{pmatrix} I_s + E_1 & E_2 \\ E_3 & E_4 \end{pmatrix},$$

with $\|E_1\|$, $\|E_2\|$, $\|E_3\|$, and $\|E_4\| \leq O(\delta)$, $rank(Z_*) = s$, $0 < s < r$, and $S \in \mathbb{F}^{r\times r}$.

Proof. Since $U = V_zR + \widetilde{V}_zS$, we have $R = V_x^*U$. By Lemma 5.4.1, for $Z \in \mathcal{B}(\mathcal{S}_\#, \delta)$, Z can be decomposed as

$$Z = \begin{pmatrix} B, \widetilde{B} \end{pmatrix} \begin{pmatrix} D_1 + \Delta_1 & 0 \\ 0 & \Delta_2 \end{pmatrix} \begin{pmatrix} B^* \\ \widetilde{B}^* \end{pmatrix}, \quad \text{i.e.,} \quad V_z = \begin{pmatrix} B, \widetilde{B} \end{pmatrix},$$

and there exists a block form $U = (U_1, U_2)$ and a $\widetilde{U} \perp (U_1, U_2)$ such that $\|P_B - P_{U_1}\| < O(\delta)$ and $\|P_{\widetilde{B}} - P_{\widetilde{U}}\| < O(\delta)$.

We now look at the block form of R:

$$R = V_z^*U = \begin{pmatrix} B^* \\ \widetilde{B}^* \end{pmatrix} (U_1, U_2) = \begin{pmatrix} B^*U_1 & B^*U_2 \\ \widetilde{B}^*U_1 & \widetilde{B}^*U_2 \end{pmatrix}.$$

Let $R = \begin{pmatrix} I_s + E_1 & E_2 \\ E_3 & E_4 \end{pmatrix}$, then $E_1 = B^* U_1 - I_s$, $E_2 = B^* U_2$, $E_3 = \widetilde{B}^* U_1$, $E_4 = \widetilde{B}^* U_2$.

For E_2, we have that

$$\begin{aligned} \|E_2\|_F^2 &= \|B^* U_2\|_F^2 = \operatorname{tr}(B^* U_2 U_2^* B) = \operatorname{tr}(B B^* U_2 U_2^*) \\ &= \operatorname{tr}(P_B P_{U_2}) = \operatorname{tr}(P_B^2 P_{U_2}^2) = \|P_B P_{U_2}\|_F^2 \leq O(\delta^2). \end{aligned}$$

Thus $\|E_2\|_F \leq O(\delta)$. Similar results hold for E_3 and E_4. As for E_1, we have

$$\|E_1\|_F^2 = \|B^* U_1\|_F^2.$$

Note that if X has distinct eigenvalues and δ is small enough, then the eigen perturbation result (Lemma 2.4.2) applies to each eigenvector. In other words, for $i = 1, \ldots, s$,

$$\|P_{B(:,i)} - P_{U_1(:,i)}\|_F \leq O(\delta).$$

Thus it can be shown that

$$\|E_1\|_F^2 = \|B^* U_1\|_F^2 = \sum_{1 \leq i,j \leq s} |B(:,i)^* U_1(:,j)|^2 \leq O(\delta^2).$$

As a result, $\|E_1\|_F \leq O(\delta)$. □

Remark 5.4.4. In the case where the eigenvalues of X are not distinct, one can simply let U_1 be the basis of the best subspace that V_z can capture, and the above result still holds.

Convergence rate outside the spurious regions

We now show that as long as the spurious regions are excluded, we can establish the linear convergence rate of the RGD, using Conditions (D) and (L) in Theorem 5.3.1.

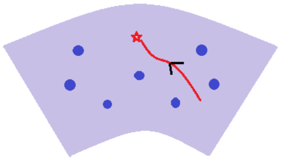


Figure 5.2: Convergence guarantee on the main part of $\mathcal{M}_r$.

Lemma 5.4.5. *Let $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$, $X, Z \in \mathcal{M}_r$. Let $\{Z_k\}_{k=0}$ be the sequence generated by the RGD with a small enough constant step size $\alpha = O(1)$. Assume that $\{Z_k\}_{k=0}$ remains in a bounded subset of $\mathcal{M}_r$, and stay in the set $\{Z : \|P_{T_z}(Z - X)\|_F \geq C_L\} \cup \{Z : \|Z - X\|_F \leq \frac{d_r}{2}\}$. Then we have:*

1) *$\|P_{T_z}(\nabla f(Z_k))\|_F \geq C_1\|Z_k - X\|_F$, for all k, with $C_1 \geq \Omega(C_L) > 0$. That is, Condition* (L) *holds with $\omega = \frac{1}{2}$;*

2) *There exists some absolute constant $C_2 > 0$ such that*

$$f_k - f_{k+1} \geq C_2\|P_{T_{Z_k}}(\nabla f(Z_k))\|_F\|Z_{k+1} - Z_k\|_F.$$

Thus, by Theorem 5.3.1, there exists $\alpha > 0$ such that the sequence $\{Z_k\}$ generated by the RGD (2.6) *converges to X in a linear convergence rate:*

$$\|Z_k - X\|_F \leq e^{-ck}.$$

Here, $c = -\log(1 - \alpha \cdot \min(\Omega(C_L^2), \frac{1}{\kappa^2}))$.

Proof. The proof is a direct consequence of Theorem 5.3.1 and Lemma 5.5.6. Specifically, to make use of Theorem 5.3.1, it suffices to check Condition (L) with $\omega = \frac{1}{2}$ and Condition (D).

As stated at the beginning of Section 5.3, we assume $\{Z_k\}$ is bounded, i.e., $\|Z\|_F \leq C$. In the region $\{Z : \|P_{T_z}(Z - X)\| \geq C_L\}$, we have $\frac{\|P_{T_z}(Z-X)\|_F}{\|Z-X\|_F} \geq \frac{C_L}{C+\|X\|_F}$; on the other hand, in the region $\{Z : \|Z - X\|_F \leq \frac{d_r}{2}\}$, by Lemma 2.4.2 and Lemma 5.5.6, $\frac{\|P_{T_z}(Z-X)\|_F^2}{\|Z-X\|_F^2} \geq \frac{d_r^2}{d_r^2+4\|X\|_F^2}$. One can take $C_1 = \min\left\{\frac{C_L}{C+\|X\|_F}, \sqrt{\frac{d_r^2}{d_r^2+4\|X\|_F^2}}\right\} > 0$. In other words, in the region $\{Z : \|P_{T_z}(Z - X)\| \geq C_L\} \cup \{Z : \|Z - X\|_F \leq \frac{d_r}{2}\}$, for $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$, we have Condition (L) holds with $\omega = \frac{1}{2}$ and $C_l = \max\left\{\frac{C+\|X\|_F}{C_L}, \sqrt{1 + \frac{4\|X\|_F^2}{d_r^2}}\right\}$.

For Condition (D), we consider

$$\begin{aligned}
f_k - f_{k+1} &= \frac{1}{2}\|Z_k - X\|_F^2 - \frac{1}{2}\|Z_{k+1} - X\|_F^2 \\
&= \frac{1}{2}\mathrm{tr}((Z_k + Z_{k+1} - 2X)(Z_k - Z_{k+1})) \\
&= \mathrm{tr}((X - Z_k)(Z_{k+1} - Z_k)) - \frac{1}{2}\|Z_{k+1} - Z_k\|_F^2.
\end{aligned}$$

Let $Z_{k+1} = Z_k + \alpha\widetilde{\xi}_k$ and $\xi_n = -P_{T_{Z_k}}(\nabla f(Z_k))$, then by first-order retraction property we have $\widetilde{\xi}_k = \xi_k + o(\|\xi_k\|_F)$. So the left- and right-hand side of Condition (D) are

$$\begin{aligned}
\text{LHS} &= f_k - f_{k+1} \\
&= \text{tr}(-\nabla f(Z_k)\alpha\widetilde{\xi}_k) - \frac{\alpha^2}{2}\|\widetilde{\xi}_k\|_F^2 \\
&= \text{tr}(-\nabla f(Z_k)(-\alpha P_{T_{Z_k}}(\nabla f(Z_k)) + o(\alpha\|\xi_k\|_F))) - \frac{\alpha^2}{2}\|\widetilde{\xi}_k\|_F^2 \\
&= \alpha\|\xi_k\|_F^2 + o(\alpha\|\xi_k\|_F), \\
\text{RHS} &= C_d\|P_{T_{Z_k}}(\nabla f(Z_k))\|_F\|Z_{k+1} - Z_k\|_F \\
&= C_d\|\xi_k\|_F\|\alpha\widetilde{\xi}_k\|_F \\
&= C_d\alpha\|\xi_k\|_F^2 + o(\alpha\|\xi_k\|_F^2).
\end{aligned}$$

By choosing a proper $C_d > 0$ and a small enough step size α, one can get

$$f_k - f_{k+1} \geq C_d\|P_{T_{z_k}}(\nabla f(Z_k))\|_F\|X_{k+1} - X_k\|_F,$$

which is Condition (D). The results now follow from Theorem 5.3.1. And the constant in the exponent for the linear convergence rate is $c = -\log(1 - \Omega(\frac{\alpha}{C_l^2})) = -\log(1 - \alpha \cdot \min(\Omega(C_L^2), \frac{1}{\kappa^2}))$. □

Lemma 5.4.5 ensures linear convergence to X in the main part of the manifold outside of the spurious regions. The rest of the analysis thus evolves around how the trajectory of the Riemannian gradient descent escapes the spurious regions.

5.5 Dynamics of the trajectory behavior

In Section 5.2, we have outlined how the trajectory of the Riemannian gradient descent is divided into three stages, corresponding to Theorem 5.2.1, Theorem 5.2.2, and Theorem 5.2.3. In this section, we introduce a few technical lemmas describing the dynamics of the trajectory behavior, which will be used for the proofs of Theorems 5.2.1–5.2.3.

We first define the general random distribution on the low-rank manifold.

Definition 5.5.1 (General random distribution)**.** Z is said to be drawn from a *general random initialization*, if $Z = V_1\Sigma V_2^*$ where V_1 and V_2 are drawn from a uniform distribution on the Stiefel manifold $\mathbb{V}_r(\mathbb{R}^n)$, and the entries of Σ

are drawn independently from a uniform distribution over $[C_1, C_2]$ with $C_2 > C_1 \geq 0$.

Remark 5.5.2. 1) The simplest example of a general random distribution is the following symmetric rank-1 Gaussian sampling. One can construct $Z_0 = cu_0u_0^* \in \mathbb{F}^{n\times n}$, where $c > 0$ is a constant and u_0 is drawn from $\frac{1}{\sqrt{n}}\mathcal{N}(0,1)^n$ for $\mathbb{F} = \mathbb{R}$, or $\frac{1}{\sqrt{2n}}\mathcal{N}(0,1)^n + i\frac{1}{\sqrt{2n}}\mathcal{N}(0,1)^n$ for $\mathbb{F} = \mathbb{C}$. It is equivalent to $Z_0 = \rho v_0 v_0^*$ with v_0 drawn from $\mathbb{V}_1(\mathbb{R}^n)$ or $\mathbb{V}_1(\mathbb{C}^n)$ and ρ drawn from $\Omega(\frac{1}{n})\chi^2(n)$ or $\Omega(\frac{1}{2n})\chi^2(2n)$. Here, $\chi^2(s) := \sum_{i=1}^{s} n_i^2$ where n_i, $i = 1, 2, ..., s$ are i.i.d. standard normal variables.

2) In practical computation, a uniform distribution on the Stiefel manifold can be easily constructed as follows. For $\mathbb{F} = \mathbb{R}$, let G be drawn from $\frac{1}{\sqrt{n}}\mathcal{N}(0,1)^{n\times r}$, and construct $V = G(G^*G)^{-\frac{1}{2}}$. For $\mathbb{F} = \mathbb{C}$, change the law of G to $\frac{1}{\sqrt{2n}}(\mathcal{N}(0,1) + i \cdot \mathcal{N}(0,1))^{n\times r}$.

3) The density functions of the sampling laws of V and Σ can be extended from constants to more general ones. In fact, it suffices to require that $\rho(V) \in [c, C]$ for any $V \in \mathbb{V}_r(\mathbb{R}^n)$, where $C > c > 0$. This allows more flexibility in the initialization.

For any given $i \in [r]$, the marginal distribution of $V_i := V(:, i)$ is a uniform distribution on $\mathbb{S}^{n-1}$. Therefore, for any given u that satisfies $\|u\|_2 = 1$, we have $\|V_i^* u\| \gtrsim \frac{1}{\text{poly}(n)}$ with high probability. More specifically, we have the following lemma.

Lemma 5.5.3. *Assume $W = (W_1, W_2, ..., W_r) \sim$ Uniform $(V_r(\mathbb{R}^n))$. For any $u_0 \in \mathbb{F}^n$ such that $\|u_0\|_2 = 1$, we have:*

1) $\mathbb{E}(\|u_0^* W\|_2^2) = \frac{r}{n}$;

2) $Prob(|u_0^* W_i|^2 \geq \Omega(\frac{1}{n \log^p n})) \geq 1 - e^{-\Omega(n)} - O(\frac{1}{\log^p n})$, $i \in [r]$;

3) $Prob(|u_0^* W_i|^2 \geq \Omega(\frac{\log n}{n})) \leq \frac{1}{poly(n)} + e^{-\Omega(n)}$, $i \in [r]$;

4) $Prob(|u_0^* W_i|^2 \geq \Omega(\frac{1}{n^{p+1}})) \geq 1 - e^{-\Omega(n)} - O(\frac{1}{n^p})$, $i \in [r]$.

Proof. Since the marginal distribution of $W(:, i)$ (i = 1, 2, . . . , r) is the uniform distribution on $\mathbb{S}^{n-1}$ and the distribution of W is right-rotational invariant, we assume unitary $U \in \mathbb{F}^{n\times n}$ such that $Uu_0 = e_1$ and $UW = \widetilde{W}$ and use the distribution of $\widetilde{W}$ to replace that of W. We first look at the real case $\mathbb{F} = \mathbb{R}$. Consider the marginal distribution of W_i, and write $W_i = \frac{g}{\|g\|}$ where g is drawn from $\mathcal{N}(0, I_n)$. Then, by the Bernstein-type inequality, we have the following estimation:

$$\text{Prob}\left(\left|\|g\|^2 - \mathbb{E}(\|g\|^2)\right| > t\mathbb{E}(\|g\|^2)\right) \leq 2\exp\left(-c\min\left(\frac{t^2\mathbb{E}(\|g\|^2)^2}{K^2 n}, \frac{t\mathbb{E}(\|g\|^2)}{K}\right)\right).$$

This implies $\text{Prob}(\|g\|^2 \in (\frac{1}{2}n, \frac{3}{2}n)) \geq 1 - e^{-\Omega(n)}$. On the other hand, we have

$$\begin{aligned} \text{Prob}(|g(1)|^2 - 1 > t) &= \text{Prob}(|g(1)| > \sqrt{1+t}) \\ &= 2\int_{\sqrt{1+t}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx =: \sqrt{\frac{2}{\pi}} A, \\ \text{where} \qquad A^2 &= \int_{\sqrt{1+t}}^{\infty}\int_{\sqrt{1+t}}^{\infty} e^{-\frac{x^2+y^2}{2}}\, dxdy = \frac{\pi}{2} e^{-\frac{1+t}{2}}. \end{aligned}$$

Taking $t = -1 + c_1 \frac{1}{\log^p n}$, we have

$$\text{Prob}\left(|g(1)|^2 > c_1 \frac{1}{\log^p n}\right) = e^{-c_1/(4\log^p n)} \geq 1 - O(\frac{1}{\log^p n}).$$

Therefore, we have $|u_0^* W_i|^2 \gtrsim \frac{1}{n\log n}$ with the fail probability controlled by $e^{-\Omega(n)} + O(\frac{1}{\log n})$. By taking $t = c_1 \log n - 1$, with $c_1 = \Omega(1)$, we have:

$$\text{Prob}(|g(1)|^2 \geq c_1 \log n) \leq e^{-c_1 \log(n)/4} = \frac{1}{n^{c_1/4}}.$$

Therefore, we have $|u_0^* W_i|^2 \lesssim \frac{\log n}{n}$ holds, with the fail probability controlled by $e^{-\Omega(n)} + \frac{1}{\text{poly}(n)}$. Taking $t = -1 + c_2/n^p$, we have that

$$\frac{|g(1)|^2}{\|g\|^2} \gtrsim \frac{1}{n^{p+1}}$$

holds with fail probability controlled by $e^{-\Omega(n)} + O(\frac{1}{n^p})$. For the complex case, the proof is very similar and we omit the details here. □

Lemma 5.5.3 shows that the general random distribution with high probability captures weak information of a given column space. The following lemma is an important technical result that describes the dynamics of the eigenvalues and column spaces of the iterative sequence.

Lemma 5.5.4. *Consider the gradient flow of $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$. Assume $Z = V_z\Sigma_z V_z^*$ and $X = UDU^*$ are the eigenvalue decompositions of the SPSD/HPSD matrices Z and X respectively. Denote $U = V_zR + \widetilde{V}_zQ$ with $\widetilde{V}_z \in Col(V_z)^{\perp}$. We have:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{r}\|D^{\frac{1}{2}}R^*e_i\|^2 = \sum_{i=1}^{r}\frac{2}{\Sigma_z(i,i)}\|QDR^*e_i\|^2.$$

Furthermore, denote the spectra of R^*R and RR^* by

$$\Sigma_{RR} = diag\{r_1, r_2, ..., r_r\}, \quad r_1 \geq \ldots \geq r_r, \tag{5.3}$$

then we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\sum_{i=1}^{r} r_i) \gtrsim \sum_{i=1}^{r}(1 - r_i)r_i, \tag{5.4}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Sigma_z(i,i)) = (RDR^*)(i,i) - \Sigma_z(i,i). \tag{5.5}$$

Moreover, $\Omega(\frac{1}{\sigma_1}\sum_{i=1}^{r}(1 - r_i)r_i) \leq \frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{r} r_i \leq \Omega(\frac{1}{\sigma_r}\sum_{i=1}^{r}(1 - r_i)r_i)$. If $\sigma_i = \Omega(1)$ for all $i \in [r]$, then we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{r} r_i = \Omega(\sum_{i=1}^{r}(1 - r_i)r_i). \tag{5.6}$$

Proof. 1) Let $Z = U_zS_zU_z^*$ be an alternative decomposition of Z with $U_z^*U_z = I_r$ and $S_z \in \mathbb{R}^{r\times r}$. Assume $P \in \mathbb{O}(r)$ such that $V_z = U_zP$ and $P^*S_zP = \Sigma_z$. Denote $U = U_zR_\sharp + \widetilde{U}_zQ_\sharp = V_zR + \widetilde{V}_zQ$. By applying the dynamical low-rank approximation from [85] (see also Lemma 4.2.2), we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}U_z = \widetilde{U}_zQ_\sharp DR_\sharp^*S_z^{-1}$$
$$\frac{\mathrm{d}}{\mathrm{d}t}S_z = R_\sharp DR_\sharp^* - S_z.$$

Therefore, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}R_\sharp = \frac{\mathrm{d}}{\mathrm{d}t}(U_z^*U) = S_z^{-1}R_\sharp DQ_\sharp^*Q_\sharp.$$

Using $R = V_z^*U = P^*U_z^*U = P^*R_\sharp$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}R &= \frac{\mathrm{d}}{\mathrm{d}t}(P^*R_\sharp)\\ &= \frac{\mathrm{d}}{\mathrm{d}t}(P)^*R_\sharp + P^*\frac{\mathrm{d}}{\mathrm{d}t}(R_\sharp)\\ &= \frac{\mathrm{d}}{\mathrm{d}t}(P)^*PR + P^*S_z^{-1}PRDQ_\sharp^*Q_\sharp\\ &= \left(\frac{\mathrm{d}}{\mathrm{d}t}(P)^*P\right)R + \Sigma_z^{-1}RDQ^*Q.\end{aligned} \tag{5.7}$$

The last equation follows from $P^*S_z^{-1}P = \Sigma_z^{-1}$ and $Q^*Q = I_r - R^*R = I_r - R^*P^*PR = Q_\sharp^* Q_\sharp$. Then, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(RDR^*) &= \Sigma_z^{-1}RDQ^*QDR^* + RDQ^*QDR^*\Sigma_z^{-1} \\ &\quad + \left(\frac{\mathrm{d}}{\mathrm{d}t}(P)^*P\right)RDR^* + RDR^*\left(P^*\frac{\mathrm{d}}{\mathrm{d}t}(P)\right).\end{aligned} \tag{5.8}$$

Due to the fact that $P^*P = I_r$, we have $\frac{\mathrm{d}}{\mathrm{d}t}(P)^*P + P^*\frac{\mathrm{d}}{\mathrm{d}t}(P) = 0$. Denote $M = P^*\frac{\mathrm{d}}{\mathrm{d}t}(P)$, then M is an antisymmetric matrix and satisfies $M+M^* = 0$. Therefore, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(RDR^*) = M^*RDR^* + RDR^*M + \Sigma_z^{-1}RDQ^*QDR^* + RDQ^*QDR^*\Sigma_z^{-1}.$$

On the other hand, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\Sigma_z) &= \frac{\mathrm{d}}{\mathrm{d}t}(P^*S_zP) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}(P)^*S_zP + P^*\frac{\mathrm{d}}{\mathrm{d}t}(S_z)P + P^*S_z\frac{\mathrm{d}}{\mathrm{d}t}(P) \\ &= P^*(R_\sharp DR_\sharp^* - S_z)P + \left(\frac{\mathrm{d}}{\mathrm{d}t}(P)^*P\right)\Sigma_z + \Sigma_z\left(P^*\frac{\mathrm{d}}{\mathrm{d}t}(P)\right) \\ &= RDR^* - \Sigma_z + \left(\frac{\mathrm{d}}{\mathrm{d}t}(P_z)^*P_z\right)\Sigma_z + \Sigma_z\left(P_z^*\frac{\mathrm{d}}{\mathrm{d}t}(P_z)\right).\end{aligned}$$

Thus, we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Sigma_z(s,s)) = (RDR^*)(s,s) - \Sigma_z(s,s). \tag{5.9}$$

Since $\frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{offdiag}(\Sigma_z)) = 0$, we have

$$\mathrm{offdiag}(RDR^*) = -M^*\Sigma_z - \Sigma_zM = M\Sigma_z + \Sigma_zM^*. \tag{5.10}$$

By substitution, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(RDR^*)(i,i) &= e_i^*(2M\Sigma_zM - \Sigma_zMM - MM\Sigma_z)e_i + \frac{2}{\sigma_i(Z)}\|QDR^*e_i\|_2^2 \\ &= 2\sum_{k=1}^{r}(\sigma_i(Z) - \sigma_k(Z))M(k,i)^2 + \frac{2}{\sigma_i(Z)}\|QDR^*e_i\|_2^2 \\ &= 2\sum_{k\neq i}\frac{1}{\sigma_i - \sigma_k}(RDR^*)(k,i)^2 + \frac{2}{\sigma_i(Z)}\|QDR^*e_i\|_2^2.\end{aligned} \tag{5.11}$$

Note that

$$\begin{aligned}\sum_{i=1}^{r}\sum_{k=1}^{r}(\sigma_i(Z)-\sigma_k(Z))M(k,i)^2 &= -\sum_{k=1}^{r}\sum_{i=1}^{r}(\sigma_k(Z)-\sigma_i(Z))M(k,j)^2\\ &= -\sum_{k=1}^{r}\sum_{i=1}^{r}(\sigma_i(Z)-\sigma_k(Z))M(k,i)^2.\end{aligned}$$

Thus, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(\sum_{i=1}^{r}\|D^{\frac{1}{2}}R^*e_i\|^2) = \sum_{i=1}^{r}\frac{2}{\sigma_i(Z)}\|QDR^*e_i\|^2. \tag{5.12}$$

2) Recall that we define the spectra of R^*R and RR^* as $\Sigma_{RR} = \operatorname{diag}\{r_1,\dots,r_r\}$. Let $R^*R = T_1\Sigma_{RR}T_1^*$ be the eigenvalue decomposition of R^*R. Since $R^*R+Q^*Q = I$, we have $Q^*Q = T_1\Sigma_{QQ}T_1^*$ with $\Sigma_{RR}+\Sigma_{QQ} = I_r$. Consider

$$\sum_{i=1}^{r}(RDR^*)(i,i) = \operatorname{trace}(RDR^*) = \operatorname{trace}(DR^*R) = \operatorname{trace}(T_1^*DT_1\Sigma_{RR}).$$

Define

$$t_i := \|D^{\frac{1}{2}}T_1e_i\|^2 = \Omega(1), \quad i\in[r], \tag{5.13}$$

then we have

$$\sum_{i=1}^{r}(RDR^*)(i,i) = \sum_{i=1}^{r}t_ir_i.$$

Now, for a bounded trajectory $\{Z_t\}$, we have

$$\begin{aligned}\sum_{i=1}^{r}\frac{2}{\sigma_i}\|QDR^*e_i\|^2 &\gtrsim \operatorname{trace}(RDQ^*QDR^*)\\ &= \operatorname{trace}(T_1^*DT_1\Sigma_{QQ}T_1^*DT_1\Sigma_{RR}).\end{aligned}$$

Denote $N = T_1^*DT_1$, then $N = \Omega(1)$ because D is the spectra of X and T_1 is orthonormal. We obtain the following estimate:

$$\begin{aligned}\sum_{i=1}^{r}\frac{2}{\sigma_i}\|QDR^*e_i\|^2 &\gtrsim \sum_{i=1}^{r}r_i(1-r_i)N(i,i)^2\\ &\gtrsim \sum_{i=1}^{r}r_i(1-r_i).\end{aligned}$$

This gives

$$\frac{\mathrm{d}}{\mathrm{d}t}(\sum_{i=1}^{r} r_i) \gtrsim \sum_{i=1}^{r} r_i(1-r_i).$$

More specifically, we have $\Omega\left(\frac{1}{\sigma_1}\sum_{i=1}^{r}(1-r_i)r_i\right) \le \frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{r} r_i \le \Omega\left(\frac{1}{\sigma_r}\sum_{i=1}^{r}(1-r_i)r_i\right)$. If $\sigma_i = \Omega(1)$, for all $i \in [r]$, by Equation (5.12), we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\sum_{i=1}^{r} \|D^{\frac{1}{2}}R^* e_i\|^2) &= \Omega\left(\sum_{i=1}^{r} \|QDR^* e_i\|^2\right)\\ &= \Omega(\mathrm{trace}(T_1^* D T_1 \Sigma_{QQ} T_1^* D T_1 \Sigma_{RR}))\\ &= \Omega\left(\sum_{i=1}^{r} r_i(1-r_i)\right).\end{aligned}$$

Thus we have $\frac{\mathrm{d}}{\mathrm{d}t}\sum_{j=1}^{r} r_i = \Omega\left(\sum_{i=1}^{r}(1-r_i)r_i\right)$.

□

The values of $\{r_i\}_{i=1}^{r}$ describe the angle between the column spaces of Z and X. From (5.4), we can conclude that $\sum_{i=1}^{r} r_i$ is non-decreasing. The quantity $\sum_{i=1}^{r} r_i$ can be used to describe whether Z is close to a spurious critical point.

The following lemma gives a more detailed description of the dynamics of R defined above.

Lemma 5.5.5. *Under the same setting as Lemma 5.5.4 and defining $M = P^*\frac{\mathrm{d}}{\mathrm{d}t}(P)$, we have the following:*

$$\frac{\mathrm{d}}{\mathrm{d}t}R = M^* R + \Sigma_z^{-1} RDQ^*Q, \tag{5.14}$$

$$M(j,i) = \frac{RDR^*(j,i)}{\sigma_i - \sigma_j}, \quad j \neq i, \tag{5.15}$$

$$\|Me_i\| \lesssim \frac{1}{\sigma_i}\|R^* e_i\|. \tag{5.16}$$

Proof. Equation (5.14) can be deduced from (5.7), and (5.15) can be deduced from (5.10) directly.

To prove (5.16), by (5.14), we have

$$\frac{\mathrm{d}}{\mathrm{d}t}R(i,k) = \frac{1}{\sigma_i}\cdot d_k(1-\|Re_k\|^2)\cdot R(i,k) - \frac{1}{\sigma_i}\cdot\sum_{s\neq k} d_s R(i,s)\langle Re_s, Re_k\rangle + e_i^* M^* Re_k.$$

By (5.15), $M(i',i) = \frac{RDR^*(i',i)}{\sigma_i - \sigma_{i'}}$, and we have

$$\frac{\mathrm{d}}{\mathrm{d}t}R(i,k) = \frac{1}{\sigma_i} \cdot d_k(1 - \|Re_k\|^2) \cdot R(i,k) - \frac{1}{\sigma_i} \cdot \sum_{s \neq k} d_k R(i,s)\langle Re_s, Re_k\rangle \pm O(\frac{1}{\sigma_i})\|R^* e_i\|\|Re_k\|).$$

By Assumption 2, we have

$$\|Me_i\| \lesssim \frac{1}{c_g}\|R^* e_i\| \cdot \max_{i' \in [r]\setminus\{i\}} \min(\frac{1}{\sigma_{i'}}, \frac{1}{\sigma_i})\|R^* e_{i'}\| \lesssim \frac{1}{\sigma_i}\|R^* e_i\|.$$

Therefore, $|e_i^* M^* Re_k| \leq \frac{O(1)}{\sigma_i}\|R^* e_i\|\|Re_k\|$. □

Finally, the following lemma gives the local Łojasiewicz inequality in the neighborhood of the ground truth X as well as the local convergence rate.

Lemma 5.5.6 (Local convergence)**.** *For $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$, let $Z = U_z \Sigma V_z^*$ and $X = UDV^*$. Denote σ_j and d_j as the j-th largest eigenvalue of Z and X respectively.*

1) *We have*

$$\frac{\sigma_r^2}{\sigma_r^2 + \|X\|_F^2}\|Z - X\|_F^2 \leq \|P_{T_z}(Z - X)\|_F^2 \leq \|Z - X\|_F^2;$$

2) *In the local region around the ground truth $\{Z : rank(Z) = r,\ \|Z - X\|_F < \frac{d_r}{2}\}$, we have $\sigma_r > \frac{d_r}{2} > 0$ and $\frac{\|P_{T_z}(Z-X)\|_F}{\|Z-X\|_F} \gtrsim d_r$. Furthermore, with a small enough constant step size α, we have $\|\phi_\alpha(Z) - X\|_F \leq e^{-c}\|Z - X\|_F$ with $c = -\log(1 - \Omega(\alpha d_r^2)) > 0$.*

Proof. For simplicity, the following proof is based on the symmetric case where $Z = U_z \Sigma U_z^*$ and $X = UDU^*$.

1) Let $\widetilde{U}_z \in \mathbb{F}^{n \times (n-r)}$ be the orthogonal complement of U_z, and

$$X = (U_z, \widetilde{U}_z)\widetilde{X}(U_z, \widetilde{U}_z)^* = (U_z, \widetilde{U}_z)\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}\begin{pmatrix} U_z^* \\ \widetilde{U}_z^* \end{pmatrix}.$$

Then we have

$$\|P_{T_z}(Z - X)\|_F = \left\|\begin{pmatrix} X_{11} - \Sigma_z & X_{12} \\ X_{21} & 0 \end{pmatrix}\right\|_F = \|X_{11} - \Sigma_z\|_F^2 + \|X_{12}\|_F^2 + \|X_{21}\|_F^2,$$

$$\|Z - X\|_F^2 = \|P_{T_z}(Z - X)\|_F^2 + \|X_{22}\|_F^2.$$

We also assume that $U = U_z R + \widetilde{U}_z S$, where $R \in \mathbb{F}^{r\times r}$ and $S \in \mathbb{F}^{(n-r)\times r}$. Then, we obtain

$$X_{11} = RDR^*, \quad X_{12} = RDS^*, \quad X_{21} = SDR^*, \quad X_{22} = SDS^*.$$

The goal is to find lower and upper bounds for

$$s = \frac{\|P_{T_z}(Z - X)\|_F^2}{\|Z - X\|_F^2}.$$

It is obvious that $s \leq 1$. To identify the lower bound we consider

$$\phi = \frac{\|Z - X\|_F^2}{\|X_{22}\|_F^2}.$$

Then we have $s = \frac{\phi - 1}{\phi}$. Note that $\phi \geq 1$ because $\|Z - X\|_F^2 = \|P_{T_z}(Z - X)\|_F^2 + \|X_{22}\|_F^2 \geq \|X_{22}\|_F^2$. For fixed Σ and D, ϕ can be seen as a function of R. We express $\phi(R)$ in terms of ϕ_1 and ϕ_2 defined below:

$$\phi(R) = \frac{\phi_1(R)}{\phi_2(R)}, \quad R \in \mathcal{D} \subset \mathbb{F}^{r\times r}, \ \mathcal{D} = \{R : 0 \preccurlyeq R^*R \preccurlyeq I_r\}, \tag{5.17}$$

where

$$\begin{aligned}\phi_1(R) &= \|X\|_F^2 + \|\Sigma\|_F^2 - 2\langle \Sigma, X_{11}\rangle = \|D\|_F^2 + \|\Sigma\|_F^2 - 2\mathrm{tr}(\Sigma RDR^*),\\ \phi_2(R) &= \mathrm{tr}\left((SDS^*)^2\right) = \mathrm{tr}\left((DS^*S)^2\right) = \mathrm{tr}\left((D(I - R^*R))^2\right).\end{aligned}$$

To minimize ϕ for given D and Σ, the first-order condition is obtained by taking derivative of ϕ over R:

$$\begin{aligned}\frac{\partial \phi}{\partial R} &= \frac{1}{\phi_2^2}(\frac{\partial \phi_1}{\partial R}\phi_2 - \frac{\partial \phi_2}{\partial R}\phi_1)\\ &= \frac{1}{\phi_2^2}(-4\phi_2 \Sigma RD + 4\phi_1 RD(I - R^*R)D)\\ &= \frac{4}{\phi_2^2}(-\phi_2\Sigma RD + \phi_1 RDS^*SD).\end{aligned}$$

Imposing $\frac{\partial \phi}{\partial R} = 0$ gives

$$\Sigma RD = \phi \cdot RDS^*SD. \tag{5.18}$$

We now claim that this first-order condition cannot be satisfied in the interior of the domain $\mathcal{D}$. To see this, note that the above equation (5.18) gives

$$\Sigma RDR^* = \phi \cdot RDS^*SDR^*, \quad \text{i.e., } \langle \Sigma, X_{11}\rangle = \phi\|X_{12}\|_F^2 = \phi\|X_{21}\|_F^2.$$

Thus, we have

$$\begin{aligned}\phi &= \frac{\|Z-X\|_F^2}{\|X_{22}\|_F^2} = \frac{\|\Sigma\|_F^2 + \|X_{11}\|_F^2 - 2\langle \Sigma, X_{11}\rangle + \|X_{12}\|_F^2 + \|X_{21}\|_F^2 + \|X_{22}\|_F^2}{\|X_{22}\|_F^2} \\ &= \frac{\|\Sigma\|_F^2 + \|X_{11}\|_F^2 - (2-2/\phi)\langle \Sigma, X_{11}\rangle}{\|X_{22}\|_F^2} + 1 \\ &= \frac{\|X_{11} - (1-1/\phi)\Sigma\|_F^2}{\|X_{22}\|_F^2} + (\frac{2}{\phi} - \frac{1}{\phi^2})\frac{\|\Sigma\|_F^2}{\|X_{22}\|_F^2} + 1 \\ &\geq (\frac{2}{\phi} - \frac{1}{\phi^2})\frac{\|\Sigma\|_F^2}{\|X_{22}\|_F^2} + 1.\end{aligned}$$

Note that when $R \in \operatorname{int}(\mathcal{D})$, R is full-rank. Since D, Σ are also full-rank, (5.18) gives $\phi^2 = \|\Sigma\|_F^2 / \|X_{22}\|_F^2$ by some simple matrix manipulation:

$$\begin{aligned} & \Sigma R D = \phi \cdot R D S^* S D \\ \Rightarrow\quad & R^{-1}\Sigma R = \phi \cdot D S^* S \\ \Rightarrow\quad & \operatorname{tr}\left((R^{-1}\Sigma R)^2\right) = \phi^2 \operatorname{tr}\left((DS^*S)^2\right) \\ \Rightarrow\quad & \operatorname{tr}\left(\Sigma^2\right) = \phi^2 \operatorname{tr}\left((SDS^*)^2\right) \\ \Rightarrow\quad & \|\Sigma\|_F^2 = \phi^2 \|X_{22}\|_F^2. \end{aligned}$$

Hence, we obtain

$$\phi \geq (\frac{2}{\phi} - \frac{1}{\phi^2})\phi^2 + 1 = 2\phi - 1 + 1 = 2\phi,$$

which contradicts the fact that $\phi \geq 1$.

Therefore, either the extreme values of ϕ are only achieved on $\partial\mathcal{D}$, or the full-rankness of D is violated. In either case, the RHS of (5.18) is rank-deficient. So the LHS of (5.18) (and thus R and X_{11}) is also rank-deficient. Therefore, we conclude that

$$\phi - 1 \geq \frac{\|X_{11} - \Sigma\|_F^2}{\|X_{22}\|_F^2} \geq \frac{\|X_{11} - \Sigma\|_F^2}{\|X\|_F^2} \geq \frac{\sigma_r^2}{\|X\|_F^2},$$

where σ_r is the smallest nonzero eigenvalue of Z. Finally, we obtain

$$s \geq \frac{\sigma_r^2}{\sigma_r^2 + \|X\|_F^2}.$$

Now we have $\|P_{T_z}(Z - X)\|_F^2 \geq s\|Z - X\|_F^2$ for $s > 0$ as long as Z is of rank r and the smallest nonzero eigenvalue of Z is bounded away from 0.

2) By Lemma 2.4.2, when $\|Z - X\|_F < \frac{d_r}{2}$, we have $\sigma_r > \frac{d_r}{2}$. Furthermore, by (1) we have $\frac{\|P_{T_z}(Z-X)\|_F}{\|Z-X\|_F} \gtrsim d_r$. By Lemma 5.4.5, we have the locally linear convergence toward X with $c = -\log\left(1 - \Omega(\alpha d_r^2)\right)$. To prove the result for single-step convergence, we consider

$$\begin{aligned} f_{k+1} &\leq f_k - C_2\|P_{T_z}(Z_k - X)\|_F\|Z_{k+1} - Z_k\|_F \\ &\leq f_k - C_2(\alpha + o(\alpha)) \cdot \|P_{T_z}(Z_k - X)\|_F^2 \\ &\leq (1 - \Omega(\alpha)C_1^2C_2)f_k. \end{aligned}$$

Here, $f_k := \frac{1}{2}\|Z_k - X\|_F^2$. The first inequality is from Condition (D). The second inequality is by the first-order retraction, and the third inequality is from Condition (L). Therefore, we have

$$\|Z_{k+1} - X\|_F \leq \sqrt{1 - \Omega(\alpha)C_1^2C_2}\|Z_k - X\|_F.$$

That is, $\|Z_{k+1} - X\|_F \leq e^{-c}\|Z_k - X\|_F$, with $c = -\log(1 - \Omega\left(\alpha d_r^2\right))$.

□

5.6 Proofs of Theorem 5.2.1–Theorem 5.2.3

In this section, we give the detailed proofs of Theorems 5.2.1-5.2.3 using the technical results from the previous three subsections.

Proof of Theorem 5.2.1. Our goal is to show that with high probability, the sequence starting from randomly initialized Z_0 will not reach $\cup_{Z_* \in \mathcal{S}_\# \setminus \{0\}}\mathcal{B}(Z_*, \delta)$ (the spurious regions near any other nonzero spurious critical points) within $K_0 = \Omega(1)$ steps, and will reach $\mathcal{B}(0, \delta)$ (the spurious region near 0) at step K_0.

Without loss of generality, assume $\|X\|_F = 1$. Recall that $X = UDU^*$ and $Z = V_z\Sigma_zV_z^*$ are the eigenvalue decompositions of X and Z respectively, and $U = V_zR + \widetilde{V}_zQ$. By Lemma 5.5.3, with fail probability controlled by $\frac{1}{\text{poly}(n)}$, we have $\|R\|_F \lesssim \sqrt{\frac{\log n}{n}}$. This implies $\sum_{i=1}^r r_i \lesssim \sqrt{\frac{\log n}{n}}$. By (5.6) in Lemma 5.5.4, we have $\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^r r_i = \Omega(\sum_{i=1}^r r_i(1 - r_i)) = \Omega(\sum r_i)$. Combined with Assumption 1, within $k \leq K_0 = \Omega(1)$ steps, with high probability exceeding $1 - \frac{1}{\text{poly}(n)}$, we have $\sum_{i=1}^r r_i \lesssim \sqrt{\frac{\log n}{n}} < \Omega(1)$. By Lemma 5.4.3, it implies that the sequence will not reach any spurious region $\mathcal{B}(Z_*, \delta)$ where Z_* is a nonzero spurious critical point and $\delta = \Omega(1)$.

Meanwhile, since Z_0 is drawn from the general random distribution, for all $i \in [r]$ we have $\sigma_i = \Omega(1)$, where the constant depends on C_1 and C_2 in Definition 5.5.1. By (5.5) in Lemma 5.5.4, within $\Omega(1)$ time, for every i, $\frac{\mathrm{d}}{\mathrm{d}t}\sigma_i = RDR^*(i,i) - \sigma_i = -\Omega(\sigma_i)$. Therefore, there exists $K_0 = \Omega(1)$ depending only on C_1, C_2, δ, such that within K_0 steps, $\sigma_i = \Omega(1)$ for all $i \in [r]$. After K_0 steps, we have $\sum_i \sigma_i = e^{-\Omega(K_0)} \cdot \Omega(1) \leq \delta$ and $\sum_i r_i < \Omega(\sqrt{\frac{\log n}{n}}) < \delta = \Omega(1)$. By Lemma 5.4.3, the sequence enters $\mathcal{B}(0,\delta)$. □

Proof of Theorem 5.2.2. From the proof of the previous stage, at $k = K_0$, Z_k is in the spurious region $\mathcal{B}(0,\delta)$ and $\sum_i r_i < \Omega(\sqrt{\frac{\log n}{n}})$. Moreover, by Lemma 5.5.3, with fail probability controlled by $\frac{1}{\mathrm{poly}(n)}$, we have $\|R\|_F \gtrsim \frac{1}{\mathrm{poly}(n)}$. Thus, $\frac{1}{\mathrm{poly}(n)} \lesssim \sum_i r_i \lesssim \sqrt{\frac{\log n}{n}}$. By (5.5) in Lemma 5.5.4, we have $\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^r r_i \gtrsim \sum_{i=1}^r r_i$. This implies $\sum_{i=1}^r r_i$ increases at least exponentially fast. Therefore, $\sum_{i=1}^r r_i \gtrsim \delta = \Omega(1)$ at step $K_1' = K_0 + C_1' \log n$ with some $C_1' = O(1) > 0$. By Lemma 5.4.3, it implies that the sequence has escaped from $\mathcal{B}(0,\delta)$. Since $\sum_{i=1}^r r_i$ is non-decreasing, $\sum_{i=1}^r r_i \gtrsim \delta$ for all $k \geq K_1'$. This implies the sequence will not come back to $\mathcal{B}(0,\delta)$, i.e., $\cup_{k \geq K_1'}^{\infty} Z_k \cap \mathcal{B}(0,\delta) = \emptyset$. □

Proof of Theorem 5.2.3. Under the setting of Theorem 5.1.3 and following Theorem 5.2.2, we consider the case that in stage 3, the sequence enters spurious region $\mathcal{B}(Z_*,\delta)$ near a spurious critical point Z_*. We now look at how the iterative sequence escapes from this spurious region by looking at the change of R from one step to the next. We introduce a few shorthand notations:

$$\bar{d}_i := d_i(1 - \|Re_i\|^2), \tag{5.19}$$

$$\|Re_j\|_{\backslash j} := \sqrt{\sum_{l \neq j} R(l,j)^2}, \tag{5.20}$$

$$\|R^* e_j\| := \sqrt{\sum_{l} R(j,l)^2}. \tag{5.21}$$

Now, we consider the behavior of $R(j,j)$ and $R(i,j)$. By (5.14), (5.16), and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}R(j,j) &= \frac{1}{\sigma_j}\cdot d_j(1-\|Re_j\|^2)\cdot R(j,j) - \frac{1}{\sigma_j}\cdot\sum_{k\neq j} d_k R(j,k)\langle Re_k, Re_j\rangle + e_j^* M^* Re_j \\ &= \frac{1}{\sigma_j}\cdot d_j(1-\|Re_j\|^2)\cdot R(j,j) - \frac{1}{\sigma_j}\cdot\sum_{k\neq j} d_k R(j,k)\langle Re_k, Re_j\rangle \pm O(\frac{1}{\sigma_j})\|R^* e_j\|\|Re_j\|).\end{aligned}$$

Similarly,

$$\frac{\mathrm{d}}{\mathrm{d}t}R(i,j) = \frac{1}{\sigma_i}\cdot d_j(1-\|Re_j\|^2)\cdot R(i,j) - \frac{1}{\sigma_i}\sum_{k\neq j} d_k\cdot R(i,k)\langle Re_k, Re_j\rangle \pm \frac{1}{\sigma_i}O(\|R^* e_i\|\|Re_j\|),$$

$$\frac{\mathrm{d}}{\mathrm{d}t}R(j,i) = \frac{1}{\sigma_j}\cdot d_i(1-\|Re_i\|^2)\cdot R(j,i) - \frac{1}{\sigma_j}\sum_{k\neq i} d_k\cdot R(j,k)\langle Re_k, Re_i\rangle \pm \frac{1}{\sigma_j}O(\|R^* e_j\|\|Re_i\|).$$

The dynamics can be further simplified as

$$\frac{\mathrm{d}}{\mathrm{d}t}R(j,j) = \frac{1}{\sigma_j}\cdot d_j(1-\|Re_j\|^2)\cdot R(j,j) \pm \frac{1}{\sigma_j}O(\|R^* e_j\|\|Re_j\|), \tag{5.22}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}R(i,j) = \frac{1}{\sigma_i}\cdot d_j(1-\|Re_j\|^2)\cdot R(i,j) \pm \frac{1}{\sigma_i}O(\|R^* e_i\|\|Re_j\|), \tag{5.23}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}R(j,i) = \frac{1}{\sigma_j}\cdot d_i(1-\|Re_i\|^2)\cdot R(j,i) \pm \frac{1}{\sigma_j}O(\|R^* e_j\|\|Re_i\|). \tag{5.24}$$

We will make use of the dynamics equations (5.22)–(5.24) frequently in the proof.

In stage 1 and stage 2 and within $O(\log\log\log n)$ time, by (5.5) we have that all $\{\sigma_i\}$ are in the range of $[\Omega(\frac{1}{\log\log n}), \Omega(1)]$. By (5.5) in Lemma 5.5.4, $\frac{\mathrm{d}}{\mathrm{d}t}\sigma_r = RDR^*(r,r) - \sigma_r = -\Omega(\sigma_r)$. Thus within $\Omega(\log\log\log n)$ time, $\sigma_r \gtrsim \Omega(\frac{1}{\log\log n})$.

By Lemma 5.5.3, initially with high probability no less than $1 - \frac{1}{\mathrm{poly}(\log n)}$, we have that $R(j,j)$, $R(i,j)$, $R(j,i)$ are in the range of $(\frac{1}{\sqrt{n}\mathrm{poly}(\log n)}, \frac{\sqrt{\log n}}{\sqrt{n}})$. Using (5.22)–(5.24), one can show that

$$\frac{\mathrm{d}}{\mathrm{d}t}\|Re_j\|^2 \lesssim \frac{1}{\sigma_r}\|R^* e_j\|^2, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\|R^* e_j\|^2 \lesssim \frac{1}{\sigma_r}\|R^* e_j\|^2. \tag{5.25}$$

Therefore, within $\Omega(\log\log\log n)$ time, we have that $R(j,j), R(i,j), R(j,i) = O(\frac{\mathrm{poly}(\log n)}{\sqrt{n}})$. As a consequence, through all the time within $O(\log\log\log n)$, in (5.22)–(5.24), the first term dominates the right-hand side.

Meanwhile, for $j > i$ we have that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{R(j,j)}{R(i,j)} = \Omega(1)\cdot(\frac{1}{\sigma_j} - \frac{1}{\sigma_i})\cdot\bar{d}_j\cdot\frac{R(j,j)}{R(i,j)} \gtrsim \frac{c_g\bar{d}_j}{\sigma_j}\cdot\frac{R(j,j)}{R(i,j)}. \tag{5.26}$$

By Lemma 5.5.3, at initialization, with high probability exceeding $1-\frac{1}{\mathrm{poly}(\log n)}$, for $j = r$ we have $\frac{R(r,r)}{R(i,r)} \geq \frac{1}{\mathrm{poly}(\log n)}$ for all $i \neq r$. Within $O(\log\log\log n)$-time, we have $RDR^*(r,r) = \Omega(\|R^* e_r\|^2) \leq O(\frac{\mathrm{poly}(\log n)}{n})$. On the other hand, we have seen above that $\sigma_r \gtrsim \Omega(\frac{1}{\log\log n})$. Again, by (5.5) in Lemma 5.5.4, $\frac{\mathrm{d}}{\mathrm{d}t}\sigma_r = RDR^*(r,r) - \sigma_r = -\Omega(\sigma_r)$. This implies $\frac{\mathrm{d}}{\mathrm{d}t}\frac{R(r,r)}{R(i,r)} \gtrsim e^{\Omega(1)\cdot t}\frac{R(r,r)}{R(i,r)}$. Using Grönwall's inequality, from (5.26), it takes $O(\log\log\log n)$-time to get $\frac{R(r,r)}{R(i,r)} \geq \Omega(1)$.

In the following, beyond $O(\log\log\log n)$ time, using (5.22)–(5.23), we write out the exact dynamics of the ratio for $j = r$ as follows:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{R(r,r)^2}{\|Re_r\|_{\setminus r}^2} =& \frac{2}{\sigma_r}\cdot\left(\bar{d}_r\cdot\frac{R(r,r)^2}{\|Re_r\|_{\setminus r}^2} \pm O(\frac{\|R^* e_r\|\|Re_r\|}{\|Re_r\|_{\setminus r}})\cdot\frac{R(r,r)}{\|Re_r\|_{\setminus r}}\cdot\right)\\
&\pm O\left(\frac{1}{\sigma_{r-1}}\cdot\frac{R(r,r)^2}{\|Re_r\|_{\setminus r}^2}\cdot\frac{\|Re_r\|}{\|Re_r\|_{\setminus r}}\right)\\
=&\frac{\Omega(1)}{\sigma_r}\cdot\frac{R(r,r)^2}{\|Re_r\|_{\setminus r}^2} \pm O(1)\cdot\frac{1}{\sigma_{r-1}}\cdot\frac{R(r,r)^2}{\|Re_r\|_{\setminus r}^2}\cdot\sqrt{1+\frac{R(r,r)^2}{\|Re_r\|_{\setminus r}^2}}\\
=&\frac{R(r,r)^2}{\|Re_r\|_{\setminus r}^2}\cdot\left(\frac{\Omega(1)}{\sigma_r} \pm \frac{O(1)}{\sigma_{r-1}}\cdot\sqrt{1+\frac{R(r,r)^2}{\|Re_r\|_{\setminus r}^2}}\right).
\end{aligned}$$

The first equality holds because

$$\begin{aligned}
\left|\frac{\mathrm{d}}{\mathrm{d}t}\|Re_j\|_{\setminus j}^2\right| &= \left|2\sum_{k\neq j}R(k,j)\cdot\frac{\mathrm{d}}{\mathrm{d}t}R(k,j)\right| \leq 2\|Re_j\|_{\setminus j}\sqrt{\sum_{k\neq j}\left(\frac{\mathrm{d}}{\mathrm{d}t}R(k,j)\right)^2}\\
&\leq 2\|Re_j\|_{\setminus j}\cdot\sqrt{r-1}\cdot\max_{k\neq j}\left|\frac{\mathrm{d}}{\mathrm{d}t}R(k,j)\right|\\
&= 2\|Re_j\|_{\setminus j}\cdot\sqrt{r-1}\cdot\max_{k\neq j}\left|\frac{\bar{d}_j}{\sigma_k}R(k,j) \pm \frac{1}{\sigma_k}\cdot O(\|Re_j\|)\right|\\
&\leq O(\frac{1}{\sigma_{r-1}}\|Re_j\|\|Re_j\|_{\setminus j}).
\end{aligned}$$

By Assumption 2, $\frac{\sigma_{r-1}}{\sigma_r} \geq \frac{1}{1-c_g} = \Omega(1)$. Therefore, beyond $O(\log\log\log n)$ time, we have that $\frac{R(r,r)}{\|Re_r\|_{\setminus r}} \geq \Omega(1)$.

Using (5.22), we then have that

$$\frac{\mathrm{d}}{\mathrm{d}t}R(r,r) \gtrsim \frac{1}{\sigma_r}\cdot(d_r - \|Re_r\|^2)R(r,r) - \frac{1}{\sigma_r}\cdot O(R(r,r))\cdot\|R^* e_r\|. \tag{5.27}$$

If $\|Re_r\|$ and $\|R^*e_r\|$ are smaller than a $\Omega(1)$-constant δ', we have $\frac{\mathrm{d}}{\mathrm{d}t}R(r,r) \gtrsim \frac{1}{\sigma_r} \cdot \Omega(1)R(r,r)$. This means that within a spurious region, when σ_r is small, $R(r,r)$ grows exponentially fast. It takes at most $O(\log n)$ steps for $R(r,r)$ to grow from $\Omega(\frac{1}{\mathrm{poly}(n)})$ to $\delta_c = \Omega(1)$. By Lemma 5.4.3, this means that Z escapes the spurious regions $\mathcal{B}(\mathcal{S}_\# \setminus \{0\}, \delta)$.

Assume that at stage 3 there exists some $K = (K_1 + O(1)) \cdot \frac{1}{\alpha}$ and some $Z_* \in \mathcal{S}_\# \setminus \{0\}$ such that $Z_K \in \mathcal{B}(Z_*, \delta)$ and $Z_{k \le K} \cap \mathcal{B}(Z_*, \delta) = \emptyset$, i.e., Z enters another spurious region again. From (5.22), we have $\frac{\mathrm{d}}{\mathrm{d}t}R(r,r) \ge -\frac{O(1)}{\sigma_r}R(r,r)$. Let R_+ denote the R at the next step and ΔR denote the change in R in one step. When $\alpha \cdot \sigma_r$ is smaller than some constant c, combined with Assumption 1, we have $\Delta R(r,r) \ge -\alpha \cdot O(1)\frac{1}{\sigma_r}R(r,r) \ge -\frac{2}{3}R(r,r)$. This gives $R_+(r,r) \ge \frac{1}{3}R(r,r)$. In other works, $R(r,r)$ never decreases too fast. Thus, with high probability exceeding $1 - \frac{1}{\mathrm{poly}(n)}$, $R(r,r) \ge \frac{1}{\mathrm{poly}(n)} \cdot 1 \cdot (\frac{1}{3})^{O(1)}$, at $k = K$. When inside this spurious region, we again use $\frac{\mathrm{d}}{\mathrm{d}t}R(r,r) \gtrsim \frac{1}{\sigma_r}R(r,r)$ to deduce that it takes $O(\log n) \cdot \frac{1}{\alpha}$ steps to escape this spurious region.

Outside of spurious regions, using Łojasiewicz inequality from Lemma 5.4.5, Z converges to X exponentially fast. This implies that the sequence enters spurious regions at most $O(1)$ times. Overall, it takes $O(\log \frac{1}{\epsilon}) \cdot \frac{1}{\alpha}$ steps to reach an ϵ-accurate solution. □

5.7 Results for the weak isometry case

In this section, we present the results for the weak isometry case (1.3):

$$\min_{Z \in \mathcal{M}_r} F_2(Z) = \frac{\theta}{2}(\|Z\|_F - \|X\|_F)^2 + \|Z - X\|_F^2.$$

Global convergence for least squares on $\mathcal{M}_1$

The following result can be regarded as a special case of Theorem 5.1.3 restricted to $r = 1$. However, the analysis is simpler. We list this as one of the main results because the trajectory behavior in this result may help us understand the trajectory behavior of phase retrieval in the next subsection. Specifically, when $r = 1$, by Lemma 2.3.1, we only have one spurious critical point $\mathcal{S}_\# = \{0\}$. In addition, we can directly write out the closed form formula of gradient descent or gradient flow. We now state the result in the following theorem.

Theorem 5.7.1. *Assume X satisfies rank$(X) = 1$, $\|X\|_F = 1$. We consider $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$ with $Z \in \mathcal{M}_1$. Let $\{Z_k\}$ be the sequence generated by the Riemannian* *gradient descent (the RGD) initialized at Z_0 which is drawn from the general random distribution. Denote $Z = zz^*$ and $X = xx^*$, $h = \|Z\|_F$ and $\rho = \frac{\langle X,Z\rangle}{\|X\|_F\|Z\|_F}$. Then we have the following:*

1) *The continuous evolution dynamics of the gradient flow can be described by the following ODE system:*

$$\frac{\mathrm{d}}{\mathrm{d}t}h = -h + \rho,$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\rho = 2\frac{\rho}{h}(1-\rho).$$

   Consequently, with high probability no less than $1 - \frac{1}{poly(n)}$, the Riemannian gradient flow only converges to $Z_* = X$, and it takes $O(\log n + \log\frac{1}{\epsilon})$ time to generate an ϵ-accurate solution, i.e., to get $\|Z - X\|_F \le \epsilon\|X\|_F$.

2) *In addition, the discrete evolution dynamics can be described by the following discrete system:*

$$h_{k+1} = h_k + \alpha(-h_k + \rho_k) + O(\alpha^2),$$
$$\rho_{k+1} = \rho_k + 2\alpha\frac{\rho_k}{h_k}(1-\rho_k) + O(\alpha^2).$$

   Under Assumption 1, there exists constant stepsize $\alpha > 0$ such that with high probability no less than $1 - \frac{1}{poly(n)}$, the Riemannian gradient descent only converges to the equilibrium with $\rho^* = h^* = 1$, meaning $\lim_{k\to\infty} Z_k = X$. Moreover, it takes $O(\log\frac{1}{\epsilon} + \log n)$ iterations to generate an ϵ-accurate solution, i.e., to get $\|Z - X\|_F \le \epsilon\|X\|_F$.

Theorem 5.7.1 is in preparation for the next result on the global convergence for phase retrieval. As we will see below, the population loss function F_2 of phase retrieval and its generalization differs from F_1 in that it only satisfies a weak isometry property. A detailed proof of Theorem 5.7.1 and its connection with Theorem 5.7.2 can be found in Section 5.8. We again emphasize that the techniques for Theorem 5.7.1 cannot be directly applied to Theorem 5.1.3 because the case $r > 1$ is significantly more difficult and closed form formulas are not available.

Global convergence for the population phase retrieval problem

Isometry properties weaker than the RIP (Restricted Isometry Property) are also common in various real-world applications. An example is the phase retrieval problem, whose loss function is as follows:

$$f(Z) = \frac{1}{2}\|T(Z) - y\|_2^2 = \frac{1}{2m}\sum_{j=1}^{m}\langle A_j, Z - X\rangle^2. \tag{5.28}$$

Here, $X = xx^*$ is the ground truth, $y = T(X)$, $Z = zz^*$, and $A_j = a_j a_j^*$, where $\{a_j\}_{j=1}^m$ are i.i.d. drawn from $\mathcal{N}(0, I_n)$ (if $\mathbb{F} = \mathbb{R}$) or $\frac{1}{\sqrt{2}}(\mathcal{N}(0, I_n) + i \cdot \mathcal{N}(0, I_n))$ (if $\mathbb{F} = \mathbb{C}$). To simplify the analysis, we only establish the result for the population loss function here. We focus on studying the problem on the Riemannian manifold and revealing its connection to the rank-1 isometry case (Theorem 5.7.1). Our proof complements that of [38], which establishes a complete proof for random measurements from a different viewpoint.

Theorem 5.7.2. *For the Gaussian phase retrieval problem* (5.28)*, we have the following results:*

1) *Let $F_2(Z) := \mathbb{E}f(Z)$ be the population loss of* (5.28)*. Then, we have $F_2(Z) = \frac{\theta}{2}(\|Z\|_F - \|X\|_F)^2 + c \cdot \|Z - X\|_F^2$, where $\theta = 1$, and $c = 1$ when $\mathbb{F} = \mathbb{R}$, or $c = \frac{1}{2}$ when $\mathbb{F} = \mathbb{C}$.*

2) *Without loss of generality, consider $F_2(Z) = \|T(Z) - T(X)\|_2^2 = \frac{\theta}{2}(\|Z\|_F - \|X\|_F)^2 + \|Z - X\|_F^2$, with $0 < \theta < \Omega(1)$. Denote $Z = zz^*$, $X = xx^*$, $h = \|Z\|_F$ and $\rho = \frac{\langle X, Z\rangle}{\|X\|_F\|Z\|_F}$. Let the initial point Z_0 be sampled from the general random distribution. Then, the continuous evolution dynamics of the gradient flow can be described by the following ODE system:*

$$\frac{\mathrm{d}}{\mathrm{d}t}h = \theta - (2 + \theta)h + 2\rho,$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\rho = \frac{4\rho}{h}(1 - \rho).$$

Consequently, with high probability no less than $1 - \frac{1}{poly(n)}$, the Riemannian gradient flow only converges to $Z_* = X$, and it takes $O(\log n + \log \frac{1}{\epsilon})$ time to generate an ϵ-accurate solution, i.e., to achieve $\|Z - X\|_F \leq \epsilon\|X\|_F$.

3) *Let $\{Z_k\}$ be the sequence generated by the Riemannian gradient descent initialized at Z_0 which is drawn from the general random distribution. Then the discrete evolution dynamics can be described by the following discrete system:*

$$h_{k+1} = h_k + \alpha(\theta - (2 + \theta)h_k + 2\rho_k) + O(\alpha^2),$$
$$\rho_{k+1} = \rho_k + \alpha\frac{4\rho_k}{h_k}(1 - \rho_k) + O(\alpha^2).$$

Under Assumption 1, there exists constant stepsize $\alpha > 0$ such that with high probability no less than $1 - \frac{1}{poly(n)}$, the Riemannian gradient descent converges to the equilibrium with $\rho^* = h^* = 1$ only, meaning $\lim_{k\to\infty} Z_k = X$. Moreover, it takes $O(\log\frac{1}{\epsilon} + \log n)$ iterations to generate an ϵ-accurate solution, i.e., $\|Z - X\|_F \leq \epsilon\|X\|_F$.

The proof of Theorem 5.7.2 is built upon Theorem 5.7.1, as F_2 satisfies a weaker isometry property than that of F_1. The detailed proof is given in Section 5.8.

The proofs of Theorem 5.7.1 and Theorem 5.7.2 in Section 5.8 are much simpler than that of Theorem 5.1.3, but they also follow the idea of tracking dynamics of the trajectory. They also make use of the technical results in Sections 5.3 and 5.5.

5.8 Analysis of the weak isometry case

In this section we prove Theorem 5.7.1 and Theorem 5.7.2. Theorem 5.7.1 is a special case of Theorem 5.1.3 restricted to $r = 1$, and its proof is much simpler. Theorem 5.7.2 builds upon the previous theorem, but extends the analysis to the case of weak isometry. Finally, we provide some insights on the connections between Theorem 5.7.1 and Theorem 5.7.2.

We first introduce Theorem 5.8.1, a variant of Theorem 5.3.1, as a fundamental tool for analyzing the convergence rate for functions with weak isometry as in Theorem 5.7.2. Using this theorem, we can show that if the measurement sampling operator preserves the distances of points on the manifold to the ground truth X to some extent (indicated by C_1 and C_2 in the distance-preserving condition below), and the projection operator satisfies a similar property as before, then with this $T(\cdot)$ operator, the loss function $f(Z) = \frac{1}{2}\|T(Z) - T(X)\|_F^2$ still preserves the nice properties of the original least squares loss function $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$. As a result, the sequences generated by the Riemannian gradient descent still converge to the ground truth in a linear rate on the manifold as long as they stay outside of the spurious regions.

Theorem 5.8.1. *Assume $T : \mathcal{M}_r \to \mathbb{R}^m$ is a linear operator, and $f(Z) = \frac{1}{2}\|T(Z) - T(X)\|_2^2$. If the following conditions hold:*

1. (Distance-preserving condition) $C_1\|Z_k - X\|_F \leq \|T(Z_k) - T(X)\|_2 \leq C_2\|Z_k - X\|_F$ *and* $C_1\|Z_{k+1} - Z_k\|_F \leq \|T(Z_{k+1}) - T(Z_k)\|_2 \leq C_2\|Z_{k+1} - Z_k\|_F$, *where* $C_1, C_2 > 0$ *are uniform constants for all* k*;*
2. (Critical ratio condition) $\|Z_k - X\|_F \leq C_3\|P_{T_{Z_k}}(\nabla f(Z_k))\|_F$, *where* $C_3 > 0$ *is a positive constant for all* k*.*

Then, Conditions (D) *and* (L) *hold with* $\omega = \frac{1}{2}$*. As a consequence, by Theorem 5.3.1, there exists a small enough* $\alpha > 0$ *such that the sequence* $\{Z_k\}$ *generated by the Riemannian gradient descent converges to X in a linear rate:* $\|Z_k - X\|_F \leq e^{-ck}$*, with* $c = -\log(1 - \Omega(\frac{\alpha}{C_2^2 C_3^2}))$*.*

Proof. The proof of Theorem 5.8.1 can be reduced to proving Conditions (L) and (D) from Conditions (1) and (2) in the assumptions. Note that with the linear operator $T : \mathcal{M} \to \mathbb{R}^m$, $T(Z) = \frac{1}{\sqrt{m}}(\langle A_1, Z\rangle, \langle A_2, Z\rangle, \ldots, \langle A_m, Z\rangle)^\top$ and the loss function $f(Z) = \frac{1}{2}\|T(Z) - T(X)\|_2^2$, Conditions (L) and (D) can be formulated as follows.

1) Condition (L): Łojasiewicz gradient inequality

$$\left(\frac{1}{2m}\sum_j \langle A_j, Z - X\rangle^2\right)^{1-\omega} \leq C_l\|P_{T_z}(\frac{1}{m}\sum_j \langle A_j, Z - X\rangle A_j)\|_F$$

holds with $\omega = \frac{1}{2}$.

2) Condition (D):

$$f_k - f_{k+1} \geq C_d\|P_{T_{Z_k}}(\frac{1}{m}\sum_j \langle A_j, Z_k - X\rangle A_j)\|_F\|Z_{k+1} - Z_k\|_F.$$

The proof now goes as follows.

1) To prove Condition (L), by Conditions (1) and (2) we have

$$\begin{aligned}\|P_{T_z}(\nabla f(Z))\|_F &\geq \frac{1}{C_3}\|Z - X\|_F \\ &\geq \frac{1}{C_2C_3}\|T(Z) - T(X)\| \\ &= \frac{\sqrt{2}}{C_2C_3}\,|f(Z) - f(X)|^{\frac{1}{2}},\end{aligned}$$

which implies that (L) holds with $\omega = \frac{1}{2}$ and $C_l = \frac{\sqrt{2}C_2C_3}{2} > 0$ is an absolute constant.

2) To prove Condition (D), we first consider

$$\begin{aligned} f_k - f_{k+1} &= \frac{1}{2}\|T(Z_k) - T(X)\|_2^2 - \frac{1}{2}\|T(Z_{k+1}) - T(X)\|_2^2 \\ &= \frac{1}{2}\langle T(Z_k + Z_{k+1} - 2X), T(Z_k - Z_{k+1})\rangle \\ &= \langle T^*T(X - Z_k), Z_{k+1} - Z_k\rangle - \frac{1}{2}\|T(Z_{k+1} - Z_k)\|_F^2. \end{aligned}$$

Assume that $Z_{k+1} = Z_k + \alpha\widetilde{\xi}_k$, and $-\alpha P_{T_{Z_k}}(\nabla f(Z_k)) = \alpha\xi_k$. By first-order retraction property and Condition (1), we get

$$\begin{aligned} f_k - f_{k+1} &\geq \langle -\nabla f(Z_k), \alpha\widetilde{\xi}_k\rangle - \frac{C_2^2\alpha^2}{2}\|\widetilde{\xi}_k\|^2 \\ &= \langle -\nabla f(Z_k), \alpha\xi_k\rangle + o(\alpha\|\xi_k\|_F^2) \\ &= \langle -\nabla f(Z_k), -P_{T_{Z_k}}(\nabla f(Z_k))\rangle + o(\alpha\|\xi_k\|_F^2) \\ &= \alpha\|P_{T_{Z_k}}(\nabla f(Z_k))\|_F^2 + o(\alpha\|\xi_k\|_F^2). \end{aligned}$$

On the other hand, we also obtain

$$\begin{aligned} C_d\|P_{T_{Z_k}}(\nabla f(Z_k))\|_F\|Z_{k+1} - Z_k\|_F &= C_d\|P_{T_{Z_k}}(\nabla f(Z_k))\|_F\|\alpha\widetilde{\xi}_k\|_F \\ &= C_d\alpha\|P_{T_{Z_k}}(\nabla f(Z_k))\|_F^2 + o(\alpha\|\xi_k\|_F^2). \end{aligned}$$

By choosing $C_d > 0$ small enough, we have

$$f_k - f_{k+1} \geq C_d\|P_{T_{Z_k}}(\frac{1}{m}\sum_j \langle A_j, Z_k - X\rangle A_j)\|_F\|Z_{k+1} - Z_k\|_F,$$

i.e., Condition (D) holds.

From Theorem 5.3.1, we conclude that Riemannian gradient descent for the least squares loss function $f(Z) = \frac{1}{2}\|T(Z) - T(X)\|_2^2$ converges to its global minimum linearly. Since $\|Z - X\|_F \leq \frac{1}{C_1}\|T(Z) - T(X)\|_2$ with constant $C_1 > 0$, this implies that the sequence converge to the target point X in a linear rate. □

By throwing out a controllable failure probability, many random sensing applications potentially have such a distance-preserving property. Some examples are mentioned in Section 1.3. The RIP (Restricted Isometry Property) can also be seen as a special case of this distance-preserving condition. Instead of checking the descent inequality and the Łojasiewicz inequality in

Theorem 5.3.1, we use the above conditions as a more user-friendly version for such distance-preserving cases.

We are now ready to prove Theorems 5.7.1 and 5.7.2, using Theorem 5.8.1 and following a similar but simpler strategy compared to the one for Theorem 5.1.3.

Proof of Theorem 5.7.1. Let Z_+ denote the next iterate of RGD from Z, and $\phi_\alpha(\cdot)$ denote the iteration function with step size α, $Z_+ = \phi_\alpha(Z)$. Since Z_+ is SPSD/HPSD and rank-1, we let $Z_+ = z_+ z_+^*$. Recall that $Z = zz^*$ and $X = xx^*$, $h = \|Z\|_F$ and $\rho = \frac{\langle X, Z\rangle}{\|X\|_F \|Z\|_F}$. Let $u_z = \frac{z}{\|z\|}$ and $u_{z,+} = \frac{z_+}{\|z_+\|}$. Then $\sqrt{\rho} = \langle u_z, x\rangle$. By Lemma 4.2.2, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} h = u_z^*(X - Z)u_z = \rho - h,\\
&\frac{\mathrm{d}}{\mathrm{d}t} u_z = \frac{1}{h} \cdot (I - u_z u_z^*)(X - Z)u_z = \frac{1}{h} \cdot (\sqrt{\rho} x - \rho u_z),\\
&\frac{\mathrm{d}}{\mathrm{d}t} \rho = \frac{2}{h} \cdot \rho(1 - \rho).
\end{aligned}$$

Assume $\delta > 0$ is a small constant. Since $r = 1$, by Lemma 5.4.1, the only spurious region is $\mathcal{B}(0, \delta)$. Since Z_0 is drawn from the general random distribution, with high probability no less than $1 - \frac{1}{\mathrm{poly}(n)}$, we have $\rho|_{t=0} \geq \frac{1}{\mathrm{poly}(n)}$. Observe from the third equation above that ρ is non-decreasing, and $\frac{\mathrm{d}}{\mathrm{d}t}\rho \gtrsim \rho$ until ρ approaches 1. Thus we have $\rho \gtrsim \delta = \Omega(1)$ within $O(\log n)$ time. As ρ is non-decreasing, the Riemannian gradient flow arrives in $\mathcal{M} \setminus \mathcal{B}(0, \delta)$ and remains there. By Theorem 5.8.1 and Lemma 5.3.3, it further takes no more than $O(\log \frac{1}{\epsilon})$ time to generate an ϵ-accurate solution. Combining all the above, to generate an ϵ-accurate solution, i.e., $\|Z_k - X\|_F \leq \epsilon \|X\|_F$, it takes $O(\log \frac{1}{\epsilon} + \log n)$ time for the gradient flow.

For the Riemannian gradient descent, by Assumption 1, we have

$$\begin{aligned}
h_{k+1} &= h_k + (\alpha + o(\alpha)) \cdot (-h_k + \rho_k),\\
\rho_{k+1} &= \rho_k + 2(\alpha + o(\alpha)) \cdot \frac{\rho_k}{h_k}(1 - \rho_k).
\end{aligned}$$

Using an argument similar to the continuous case, we can prove it only takes $O(\log \frac{1}{\epsilon} + \log n)$ iterations to generate an ϵ-accurate solution, i.e., $\|Z_k - X\|_F \leq \epsilon \|X\|_F$. □

Proof of Theorem 5.7.2. Recall that $Z = zz^*$ and $X = xx^*$, $h = \|Z\|_F$ and $\rho = \frac{\langle X,Z\rangle}{\|X\|_F\|Z\|_F} \in [0,1]$. If $\mathbb{F} = \mathbb{R}$, the population loss of (5.28) is

$$\begin{aligned}\mathbb{E}f(Z) &= \frac{3}{2}\|Z\|_F^2 + \frac{3}{2}\|X\|_F^2 - \|Z\|_F\|X\|_F - 2\langle Z, X\rangle \\ &= \frac{1}{2}\left(\|Z\|_F - \|X\|_F\right)^2 + \|Z - X\|_F^2.\end{aligned}$$

Since $0 \le (\|Z\|_F - \|X\|_F)^2 \le \|Z - X\|_F^2$, we have

$$\|Z - X\|_F^2 \le F(Z) \le \frac{3}{2}\|Z - X\|_F^2.$$

If $\mathbb{F} = \mathbb{C}$, the population loss of (5.28) is

$$\begin{aligned}\mathbb{E}f(Z) &= \|Z\|_F^2 + \|X\|_F^2 - \|Z\|_F\|X\|_F - \langle Z, X\rangle \\ &= \frac{1}{2}\left(\|Z\|_F - \|X\|_F\right)^2 + \frac{1}{2}\|Z - X\|_F^2.\end{aligned}$$

And $\frac{1}{2}\|Z - X\|_F^2 \le F_2(Z) \le \|Z - X\|_F^2$.

We still let Z_+ denote the next iterate of RGD from Z. Assume $Z = zz^*$, and $Z_+ = z_+z_+^*$. Let $u_z = \frac{z}{\|z\|}$ and $u_{z,+} = \frac{z_+}{\|z_+\|}$. By Lemma 4.2.2, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}h &= u_z^*\left(2X - \left(\theta + 2 - \frac{\theta}{h}\right)Z\right)u_z = 2\rho - (2+\theta)h + \theta, \\ \frac{\mathrm{d}}{\mathrm{d}t}u_z &= (I - u_zu_z^*)\left(2X - \left(\theta + 2 - \frac{\theta}{h}\right)Z\right)u_z \cdot \frac{1}{h} = \frac{2}{h}\cdot(\sqrt{\rho}x - \rho u_z), \\ \frac{\mathrm{d}}{\mathrm{d}t}\rho &= \frac{4}{h}\cdot\rho(1-\rho).\end{aligned} \tag{5.29}$$

Note that in this case, in addition to the spurious region $\mathcal{B}(0,\delta)$, there is another region $\mathcal{B}(Z_*,\delta) := \{Z : \rho \lesssim \delta, |\theta + 2 - \frac{\theta}{h}| \lesssim \delta\}$ where the Riemannian gradient is δ-small, because the Riemannian gradient is now

$$\nabla_{\mathcal{M}}F_2(Z) = \left(\theta + 2 - \frac{\theta}{h}\right)Z - 2P_{T_z}(X).$$

Similar to Lemma 5.4.1, we have

$$\|P_{T_z}\left(\nabla_{\mathcal{M}}F_2(Z)\right)\| \le \delta \iff Z \in \mathcal{B}(0,\delta)\cup\mathcal{B}(Z_*,\delta) \Rightarrow \rho \lesssim \delta. \tag{5.30}$$

Since Z_0 is drawn from the general random distribution, with high probability no less than $1-\frac{1}{\mathrm{poly}(n)}$, we have $\rho|_{t=0} \ge \frac{1}{\mathrm{poly}(n)}$. Observe from (5.29) that

$\frac{\mathrm{d}}{\mathrm{d}t}\rho \gtrsim \rho$ when $\rho < \frac{1}{2}$ and h is bounded. The boundedness of h can be easily concluded from the first equation using $0 \leq \rho \leq 1$ and $0 < \theta < \Omega(1)$. Thus we have $\rho \gtrsim \delta = \Omega(1)$ within $O(\log n)$ time. Using the non-decreasing property of ρ and the relation (5.30), we conclude that the Riemannian gradient flow arrives in $\mathcal{M} \setminus \mathcal{B}(0, \delta) \cup \mathcal{B}(Z_*, \delta)$ and remains there. By Theorem 5.8.1 and Lemma 5.3.3, it further takes no more than $O(\log \frac{1}{\epsilon})$ time to generate an ϵ-accurate solution. Combining all the above, to generate an ϵ-accurate solution, i.e., $\|Z_k - X\|_F \leq \epsilon \|X\|_F$, it needs $O(\log \frac{1}{\epsilon} + \log n)$ time.

For the Riemannian gradient descent, by Assumption 1, we have

$$\begin{aligned}
h_{k+1} &= h_k + (\alpha + o(\alpha)) \cdot (\theta - (2 + \theta)h_k + 2\rho_k), \\
\rho_{k+1} &= \rho_k + 4(\alpha + o(\alpha)) \cdot \frac{\rho_k}{h_k}(1 - \rho_k).
\end{aligned}$$

Similar to the argument for continuous case, we can prove it only takes $O(\log \frac{1}{\epsilon} + \log n)$ iterations to generate an ϵ-accurate solution, i.e., $\|Z_k - X\|_F \leq \epsilon \|X\|_F$.

□

Comparison of Theorem 5.7.1 and Theorem 5.7.2

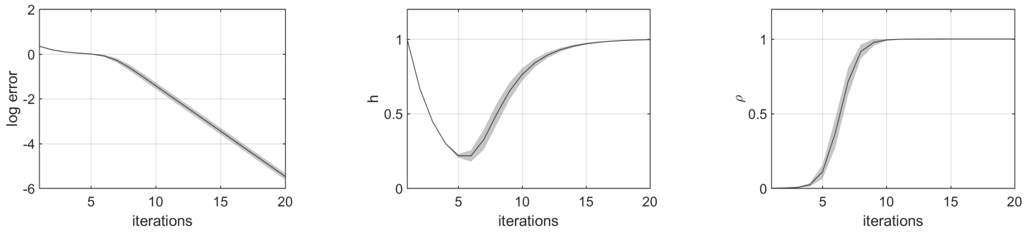
Dynamics. The dynamical low-rank approximation (Lemma 4.2.2) shows that the evolution of the column space is given by

$$\dot{U}_z = P_{U_z}^{\perp}\left(-\nabla F(Z)\right) U_z S^{-1}.$$

For $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$, we have that $\nabla F_1(Z) = Z - X$, while for $F_2(Z) = \frac{\theta}{2}(\|Z\|_F - \|X\|_F)^2 + \|Z - X\|_F^2$ we have $\nabla F_2(Z) = (2 + \theta - \frac{\theta}{h})Z - 2X$. Although $F_2(Z)$ only satisfies the weak isometry property, direct computation shows that $\dot{U}_z$ for $F_1(Z)$ and $F_2(Z)$ are similar, because $P_{U_z}^{\perp}$ on the right cancels out the Z terms and leaves only the X terms. That explains why the dynamics of $F_1(Z)$ is similar to that of $F_2(Z)$.

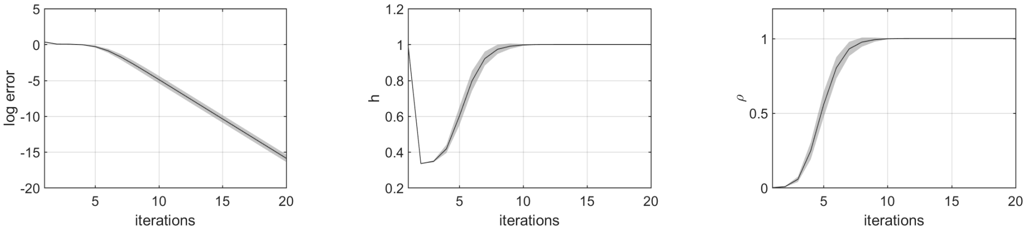
Stationary points. Theorem 5.7.1 is a special case of Theorem 5.1.3, therefore $Z_* = 0$ is the only spurious critical point and has a saddle-like geometry (see Section 5.4). On the other hand, for the phase retrieval problem in Theorem 5.7.2, it has two groups spurious critical points, which are $Z_* = 0$ and $\{Z_* : \|Z_*\|_F = \frac{\theta}{2+\theta}\|X\|_F, \langle Z, X\rangle = 0\}$. Still, the upper bound for the number of iterations that the sequence is trapped by the spurious region can be estimated in a similar way. This can be seen by comparing the proofs of Theorem 5.7.1 and Theorem 5.7.2.

Numerical illustration. To demonstrate the similarity between the evolution behavior in solving the rank-1 matrix recovery and the phase retrieval problem, we give some numerical experiments in Figure 5.3 and Figure 5.4 for a comparison. We can see that the curves of the evolution of h and ρ have similar shapes in both problems.



(a) *Log-error* (b) *The evolution of h* (c) *The evolution of ρ*

Figure 5.3: Solving the rank-1 matrix recovery by the randomly initialized Riemannian gradient descent (RGD), with $n = 1024$, $\alpha = \frac{1}{3}$. Each band stands for the results from 100 independent experiments.



(a) *Log-error* (b) *The evolution of h* (c) *The evolution of ρ*

Figure 5.4: Solving the population phase retrieval problem by the randomly initialized Riemannian gradient descent (RGD), with $n = 1024$, $\alpha = \frac{1}{3}$. Each band stands for the results from 100 independent experiments.

5.9 Discussion

In this chapter, we have established a unified framework for the analysis of a class of low-rank matrix recovery problems. We have shown that using the Riemannian gradient descent (RGD) algorithm on the low-rank matrix manifold, there is a rigorous theoretical guarantee for the fast convergence rate in low-rank matrix recovery problems.

For this purpose, we first performed an extensive analysis of the low-rank matrix manifold $\mathcal{M}_r$ itself by analyzing the simple least squares loss function $F_1(Z) = \frac{1}{2}\|Z - X\|_F^2$ where X is the ground truth. Our focus is on the symmetric positive semi-definite (SPSD) or Hermitian positive semi-definite (HPSD) setting, which is common in practice. Our results on the rank-r manifold with $r > 1$ are original and much more complicated than the corresponding results for the rank-1 case.

We showed that there is a ground truth and several spurious critical points on the manifold. The spurious critical points are of independent interest themselves, as they behave like strict saddle points, but their Hessian has singular eigen directions. In this chapter, we proved that the gradient descent or gradient flow starting from an initial guess drawn from the general random distribution converges to the ground truth with high probability. The initializations that might lead to the spurious critical points only have a small probability measure on the manifold. This result and the almost-sure escape result in Chapter 4 complement each other.

The convergence rate of the Riemannian gradient descent toward the ground truth is nearly linear and is essentially independent of the dimensionality of the problem. The main difficulty when analyzing the convergence rate comes from estimating the upper bound for the number of iterations that the sequence is trapped by the spurious regions. Our primary tool is the iteration function of the column space derived from the dynamical low-rank approximation. We showed that with high probability, the square of the initial angle between the column spaces of the ground truth and the random initialization point is between $\Omega(\frac{1}{n\log^p n})$ and $\Omega(\frac{\log n}{n})$. Moreover, by analyzing the dynamics of the trajectory, we showed that the angle grows fast in spurious regions. Thus, we showed that with high probability, the sequence generated by the randomly initialized Riemannian gradient descent escapes from the spurious regions quickly. When the sequence is outside of the spurious regions, we use the Łojasiewicz inequality tool to derive linear convergence.

The above analysis offers a general framework for a class of inverse problems that share a desirable structure, namely those problems whose forward problem is a linear mapping from a low-rank matrix to a vector and preserves the isometry property to some extent. The well-known RIP ensemble is a special case, but other applications with only weak isometry properties also fall into this category. We analyzed the phase retrieval problem as an example of weak isometry problems, and established its nearly optimal (nearly linear) convergence rate by invoking its connection with the rank-1 simple least squares problem.

The global analysis for population loss functions could potentially be extended to finite-sample problems. The finite-sample loss function is the sum of the population loss function plus some small deviations. One can control the magnitude of the deviations and show that the loss function still satisfies some weak isometry conditions. Thus the fundamental convergence guarantee by the Łojasiewicz inequality is readily applicable on the main part of the manifold. On the other hand, the geometry of the spurious regions, as well as the escape from these spurious regions, could be different from the population case. We leave the detailed analysis of finite-sample cases to future work.

Chapter 6

ROBUST LOW-RANK MATRIX RECOVERY BY RIEMANNIAN SUBGRADIENT METHOD

In this chapter, we discuss an application on the low-rank matrix manifold $\mathcal{M}_r$ that is different from the previous random linear measurements framework. The problem of interest is Robust Principal Component Analysis (RPCA), whose goal is to recover a low-rank L_* and a sparse S_* from their sum $M = L_* + S_*$. We minimize the l_1-norm loss function (6.1) over $\mathcal{M}_r$, and we use Riemannian subgradient descent to minimize this nonsmooth function.

To establish the convergence guarantee for the Riemannian subgradient descent algorithm, we first analyze the local optimality of L_* for the l_1 loss function on the manifold. We discuss the incoherence of the tangent element at L_*. We show that when the rank $r = 1$, L_* is the local minimizer of $f(L)$ on $\mathcal{M}_r$ in a local neighborhood. We then use the sharpness and weak convexity conditions to show that starting from a spectral initialization, the Riemannian subgradient algorithm converges to L_* at a linear convergence rate.

Organization of this chapter. We have given a brief introduction of the problem in Section 1.4. The rest of this chapter is organized as follows. Section 6.1 gives some preliminary information, including the problem setting and the subgradient algorithm to be used. In Section 6.2, we discuss some related work and compare their approaches with ours. In Section 6.3, we look at incoherence in the tangent space. In Section 6.4, we show the local optimality of L_* for $f(L)$ on $\mathcal{M}_1$. In Section 6.5, we establish the linear convergence rate of the subgradient algorithm toward the ground truth. Finally, we make some concluding discussion in Section 6.6

6.1 Preliminaries

In this section, we introduce the problem setting and the subgradient algorithm to be used.

Notations

Let $\mathcal{M}_r = \{X \in \mathbb{R}^{m\times n} : \mathrm{rank}(X) = r\}$ denote the low-rank matrix manifold. Let M denote the matrix to be decomposed, $M = L_* + S_*$, where L_* and S_* are unknown low-rank and sparse ground truth factors to be recovered. Let $\Omega(\cdot)$ denote the support of a matrix, and $\Omega^* = \Omega(S_*)$ denote the support of the ground truth S_*. We use $\|\cdot\|_p$ to denote the usual p norm for vectors and operator p-norm for matrices. We use $\|\cdot\|_{l_p}$ to denote the vectorized p-norm for matrices, e.g., $\|X\|_{l_1} = \sum_{i,j}|X_{ij}|$. Let $\sigma_i(\cdot)$ denote the ith largest singular value of a matrix, and σ_i specifically denotes the ith singular value of the ground truth L_*. For an integer $s > 0$, let $[s] = \{1, 2, ..., s\}$.

Problem setting

The problem of Robust PCA can be formulated in mathematical terms as follows. Given $M = L_* + S_*$, where L_* is a low-rank matrix, S_* is a sparse matrix, and both L_* and S_* are unknown, the goal is to recover L_* and S_* from M.

It is well recognized that the problem of RPCA is only uniquely solvable when L_* and S_* satisfy some incoherence and sparsity assumptions. In this work, we adopt the following assumptions.

Assumptions 6.1.1. *Assume that we have $M = L_* + S_*$, where L_* and S_* are unknown low-rank and sparse matrices satisfying the following properties:*

(a) Rank$(L_*) = r$, and $r \ll \min\{m, n\}$. Moreover, L_* is an incoherent matrix with incoherence parameter *$\mu(L_*) \le \mu$, namely, $L_* = U\Sigma V^\top$, where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, and*

$$\|U\|_{2,\infty} := \max_{1\le i\le m} \|U(i,:)\|_2 \le \frac{\mu\sqrt{r}}{\sqrt{m}},$$
$$\|V\|_{2,\infty} := \max_{1\le i\le n} \|V(i,:)\|_2 \le \frac{\mu\sqrt{r}}{\sqrt{n}}.$$

(b) S_* is a sparse matrix in the sense that for each $i \in [m]$, $j \in [n]$, we have

$$\|S(i,:)\|_0 \le pn, \quad \|S(:,j)\|_0 \le pm.$$

Let Ω^* denote the support of S_*. Moreover, the sparsity constant p satisfies

$$p \le \frac{1}{36\mu^2 r}.$$

Remark 6.1.2. We emphasize that here we do not assume the ground truth matrices L_* and S_* are generated by any underlying random model. Thus our assumptions are weaker than the assumptions using random models (e.g., the Bernoulli random sampling model). Still, they suffice to imply rich structural properties under the l_1 minimization regime.

Subgradient algorithm

In this work, we propose solving the following minimization problem on the low-rank matrix manifold $\mathcal{M}_r$:

$$f(L) = \|M - L\|_{l_1}. \tag{6.1}$$

We propose using the Riemannian subgradient algorithm to minimize (6.1) over $\mathcal{M}_r$. The subgradient algorithm has been briefly mentioned in Section 2.2, and we provide more details here for the specific problem.

The Riemannian subgradient algorithm proceeds as follows. Let L_0 be the initialization for L, and let L_k denote the kth iterate, $\{L_k\}_{k=0} \subset \mathcal{M}_r$. At the kth step, consider the subgradient of the loss function (6.1):

$$\partial f(L_k) = -\widetilde{\text{sign}}(M - L_k).$$

Here $\widetilde{\text{sign}}(\cdot)$ is the entrywise generalized sign function:

$$\widetilde{\text{sign}}(x) = \begin{cases} -1 & \text{if } x < 0; \\ t \quad \forall\, t \in [-1, 1] & \text{if } x = 0; \\ 1 & \text{if } x > 0. \end{cases}$$

In other words, any value in $[-1, 1]$ can be chosen if the variable is 0.

The Riemannian subgradient on the manifold $\mathcal{M}_r$ is the projected subgradient in the tangent space of $\mathcal{M}_r$:

$$P_{T_{L_k}}\left(\partial f(L_k)\right) = P_{T_{L_k}}\left(\widetilde{\text{sign}}(L_k - M)\right).$$

At each iteration, the Riemannian subgradient algorithm moves one step in the direction of the Riemannian subgradient. Denote the kth step size as α_k, then the next iterate is generated as follows:

$$L_{k+1} = R\left(L_k - \alpha_k \cdot P_{T_{L_k}}\left(\widetilde{\text{sign}}(L_k - M)\right)\right). \tag{6.2}$$

To sum up, we have the Riemannian subgradient method in Algorithm 1. Similar to what has been discussed in Section 2.2 for the Riemannian gradient descent method, because the tangent space projection is of rank at most $2r$, each step of the Riemannian algorithm only involves a size $2r \times 2r$ SVD, and is fast when $r \ll \min\{m, n\}$.

Algorithm 1: Riemannian subgradient algorithm

```
Input : Sparse noise corrupted low-rank matrix observation $M$, rank
        estimation $r$, initial low-rank matrix estimation $L_0$, step size
        scheme $\alpha_k$, maximum number of iterations $K$.
for $k = 0, 1, 2, \ldots, K$ do
    $L_{k+1} = R\left(L_k - \alpha_k \cdot P_{T_{L_k}}\left(\widetilde{\mathrm{sign}}(L_k - M)\right)\right)$;
    If $f(L_k) < \epsilon$, break.
end for
Truncate $M - L_k$ to obtain a sparse matrix $\hat{S}$.
Output: Recovered low-rank matrix $\hat{L} = L_k$ and sparse noise matrix $\hat{S}$.
```

To find a good initial L_0, we use the same spectral initialization method as [131]. This method finds the initial guess for S by truncating M and keeping only the largest entries in each row and each column. More specifically, $\widetilde{S} = \mathcal{T}_p(M)$, where the truncation operator $\mathcal{T}_p(\cdot)$ is defined as

$$\mathcal{T}_p(M) = \begin{cases} M(i,j), & \text{if } |M(i,j)| \geq |M(i,:)|^{[pn]} \text{ and } |M(i,j)| \geq |M(:,j)|^{[pm]}; \\ 0, & \text{otherwise.} \end{cases}$$

Here $|M(i,:)|^{[pn]}$ denotes the (pn)th largest entry in the row $|M(i,:)|$; similarly for $|M(:,j)|^{[pm]}$. We then let $\widetilde{L} = M - \widetilde{S}$. Finally we take the best rank-r approximation $L_0 = R(\widetilde{L})$ as the initial L_0. The initialization algorithm is summarized in Algorithm 2.

Note that other local initialization methods are also used in literature, with similar construction and effect. For example, truncation by ranking can be replaced with truncation by a certain threshold ζ.

The following condition comes in handy when we analyze the convergence of the algorithm.

```
Algorithm 2: Initialization
Input : Sparse noise corrupted low-rank matrix observation $M$, matrix
        size $[m, n]$, rank estimation $r$, and truncation parameter $p$.
Compute $\text{thresh}_i$ = sort(abs($M$), 1, descend), $\text{thresh}_i \leftarrow \text{thresh}_i(pm, :)$.;
Compute $\text{thresh}_j$ = sort(abs($M$), 2, descend), $\text{thresh}_j \leftarrow \text{thresh}_j(pn, :)$.;
Compute $\bar{S}$ = (abs($M$) > $\text{thresh}_i$). * (abs($M$) > $\text{thresh}_j$). * $M$;
Compute $\bar{L} = M - \bar{S}$;
Compute the SVD of $\tilde{L}$: $[U_0, \Sigma_0, V_0]$ = lansvd($\bar{L}, r$).;
$L_0 = U_0 \Sigma_0 V_0'$.;
Output: Initialization $L_0$.
```

Condition 6.1.3. *For all $k = 0, 1, \ldots$, we always have that $(L_k - L_*)$ is incoherent in the sense that*

$$\frac{\|L_k - L_*\|_{l_1}}{\|L_k - L_*\|_F} \geq c\sqrt{mn}$$

for some uniform constant $c \in (0, 1]$ independent of k.

One way to interpret the ratio $\frac{\|\cdot\|_{l_1}}{\|\cdot\|_F}$ in Condition 6.1.3 is to take it as a necessary condition for incoherence. Namely, if a matrix $X = U\Sigma V^*$ is incoherent in terms of U and V, then each entry of X must be $O(\frac{1}{\sqrt{mn}})$, which implies that the ratio $\frac{\|X\|_{l_1}}{\|X\|_F}$ is of order $\sqrt{mn}$.

We also remark that much of the results in this chapter does not depend crucially on Condition 6.1.3. We clearly distinguish those results that need Condition 6.1.3 and those that do not in the upcoming sections. The purpose of Condition 6.1.3 is to improve some constants, which results in faster convergence rate and larger local convergence neighborhood. Numerical evidence shows that this condition is consistently satisfied in practical applications.

6.2 Related work

Below we discuss some related work and compare their approaches with ours.

Convex formulations and alternating methods for RPCA. Early approaches to the Robust PCA problem mainly rely on the nuclear norm relaxation. The work [33] first proposes the Principal Component Pursuit method, which solves the RPCA problem by minimizing the weighted combination of the nuclear norm and the l_1 norm. This nuclear norm relaxation has since become a standard heuristic. Concurrent with this work, the authors of [36] develop the theory of rank-sparsity incoherence that lays the foundation for a class of low-rank recovery problems. In [37], the authors further establish a unified performance guarantee for the convex optimization approach using the nuclear norm and the l_1 norm based on the assumption of joint incoherence. However, with the growth of the dimension of the real-world problems, convex optimization methods gradually grew out of favor due to high computational cost.

An important class of nonconvex methods for the RPCA problem is the alternating projection methods. The work [108] first proposes the alternating projection algorithm that performs low-rank projection on the L variable and hard thresholding on the S variable alternatively, and establishes the linear convergence rate of this algorithm. It is worth noting that in their assumptions, no randomness is required for the support of the sparse noise S, in contrast to [37] which requires joint incoherence. Later, the algorithm is further improved by [23] for more efficient computation. In [131], the authors propose an alternating l_2 gradient descent method for RPCA, and establish its linear convergence guarantee under spectral initialization.

Manifold algorithms for RPCA. The low-rank matrix manifold $\mathcal{M}_r$ [70, 71] has recently gained attention as a useful tool for low-rank recovery problems including RPCA. Among existing work that uses $\mathcal{M}_r$ for RPCA, the one closest to us in spirit is [26], in which they essentially solve the same problem $\min_{L\in\mathcal{M}_r} f(L) = \|M - L\|_{l_1}$ over $\mathcal{M}_r$. This formulation uses l_1 minimization to promote sparsity in the S factor. However, the work [26] differs from our work in that, to tackle nonsmoothness, they solve $\min_{X\in\mathcal{M}_r} \sum_{(i,j)\in\Omega} \sqrt{\delta^2 + (X_{ij} - M_{ij})^2}$ for successively smaller δ. For each δ, the approximate problem is solved by Riemannian conjugate gradient. Moreover, the work [26] left a few questions unanswered, including the justification for the ground truth L_* being a local minima, and the convergence rate of the algorithm. In contrast, our work establishes the local optimality of $L*$ and a linear convergence guarantee under certain conditions.

Apart from the low-rank matrix manifold, some works have formulated the RPCA problem on different manifolds, usually the Grassmannian. For example, in [69], the authors propose the so-called Grassmannian robust adaptive subspace tracking (GRASTA) method, which is an alternating minimization method on the Grassmannian. In [129], the authors essentially solve $\min_{U,V} \|P_\Omega(UV^\top - M)\|_{l_2}$, where U and V are on Grassmannians, and the mask Ω is adaptively updated at each iteration to determine the support of the outliers.

Convergence of subgradient method. The Riemannian subgradient algorithm is the counterpart of Riemannian gradient algorithm for nonsmooth functions with generalized gradients. A few works have used subgradient methods on various manifolds to solve low-rank recovery problems, including orthogonal dictionary learning [11], Gaussian phase retrieval problem [46], dual principal component pursuit [135], and robust low-rank recovery [96]. The work [96] essentially studies the Gaussian matrix sensing with a gross sparse noise, and aims to recover $UU^\top$ from $y = \mathcal{A}(UU^\top) + S$, where $\mathcal{A}$ is Gaussian ensemble, but the problem is posed on the Stiefel manifold. We stress that no previous work has studied the Riemannian subgradient method on the low-rank matrix manifold $\mathcal{M}_r$ or established a provable convergence guarantee on this manifold.

When it comes to the convergence rate, it is first proposed in [45] that the sharpness condition and weak convexity condition (Conditions (S) and (C) in Lemma 6.5.2) imply linear convergence rate. The work [45] focuses on the Euclidean algorithms, and proves the convergence rate for a number of different step size schemes, including the Polyak stepsize, the constant step length, and the geometrically decaying stepsize that we use in our work. Later, the conditions (S) and (C) are extended to manifold subgradient methods by [95], where they study the Riemannian subgradient on the Stiefel manifold. In this work, we extend these conditions to the low-rank matrix manifold $\mathcal{M}_r$, which has a very different structure compared with other manifolds.

6.3 Incoherence

In this section, we take a deep dive into the incoherence property on the low-rank matrix manifold $\mathcal{M}_r$. When working with $\mathcal{M}_r$, one often deals with the tangent space. We discuss some results on the incoherence properties in the tangent space that have not been reported in previous literature.

Theorem 6.3.1. *Under Assumptions 6.1.1, a randomly sampled matrix ξ in the* *tangent space T_{L_*} at $L_* \in \mathcal{M}_r$ is incoherent with high probability. More specifically, write ξ as*

$$\xi = \begin{pmatrix} U, & \widetilde{U} \end{pmatrix} \begin{pmatrix} M & N_1 \\ N_2 & 0 \end{pmatrix} \begin{pmatrix} V^\top \\ \widetilde{V}^\top \end{pmatrix}, \tag{6.3}$$

and let $\widetilde{U}$, $\widetilde{V}$ be sampled uniformly from the Stiefel submanifolds $St(m, r; U) = \{\widetilde{U} \in \mathbb{R}^{m \times r} : \widetilde{U}^* \widetilde{U} = I_r, \widetilde{U} \perp U\}$ and $St(n, r; V) = \{\widetilde{V} \in \mathbb{R}^{n \times r} : \widetilde{V}^\top \widetilde{V} = I_r, \widetilde{V} \perp V\}$. Then with high probability no less than $1 - cN^{-3} \log N$, where $N = \min\{m, n\}$, ξ is incoherent with incoherence parameter bounded by $C + \mu$, where c and C are absolute constants.

Proof. Assume that the SVD of ξ is $\xi = U_\xi \Sigma_\xi V_\xi^\top$. Then there exist orthonormal matrices O_1, $O_2 \in \mathbb{O}_{2r}$, such that

$$U_\xi = (U, \widetilde{U}) O_1, \quad V_\xi = (V, \widetilde{V}) O_2, \quad \Sigma_\xi = O_1^\top \begin{pmatrix} M & N_1 \\ N_2 & 0 \end{pmatrix} O_2.$$

Observe that the $(2, \infty)$-norm of U_ξ is equal to that of $(U, \widetilde{U})$. This is because

$$\|U_\xi\|_{2,\infty} = \max_{1 \le i \le m} \|U_\xi(i, :)\|_2 = \max_{1 \le i \le m} \|e_i^\top \cdot (U, \widetilde{U}) O_1\|_2 = \max_{1 \le i \le m} \|e_i^\top \cdot (U, \widetilde{U})\|_2 = \|(U, \widetilde{U})\|_{2,\infty}.$$

The same applies to V_ξ. Thus it suffices to bound the $(2, \infty)$-norms of $(U, \widetilde{U})$ and $(V, \widetilde{V})$.

Using the same trick, we can assume that we first sample $\overline{U}$ and $\overline{V}$ from the Stiefel manifold $\mathrm{St}(m, r)$ and $\mathrm{St}(n, r)$ directly and re-orthogonalize them to get $\widetilde{U}$ and $\widetilde{V}$. So we only need to bound the $(2, \infty)$-norms of $(U, \overline{U})$ and $(V, \overline{V})$ because they are still the same.

Since U and V are incoherent with incoherence parameter μ by Assumptions 6.1.1, we focus on $\widetilde{U}$ and $\widetilde{V}$. Following the same idea as in the proof of [31, Lemma 2.2], one can show that there exist absolute constants c and C, such that with probability no less than $1 - cm^{-3} \log m$, we have

$$\max_{1 \le i \le m} \|P_{\overline{U}} e_i\|_2^2 \le C^2 r / m.$$

This implies that

$$\begin{aligned}\|U_\xi\|_{2,\infty} &= \|(U,\overline{U})\|_{2,\infty} = \max_{1\le i\le m} \|(U(i,:),\ \overline{U}(i,:))\|_2 \\ &\le \max_{1\le i\le m} \sqrt{\|(U(i,:)\|_2^2 + \|(\overline{U}(i,:)\|_2^2} \\ &\le \sqrt{\max_{1\le i\le m} \|(U(i,:)\|_2^2 + \max_{1\le i\le m} \|(\overline{U}(i,:)\|_2^2} \\ &\le \sqrt{\mu^2 r/m + C^2 r/m} \\ &= (\mu + C)\sqrt{\frac{r}{m}}.\end{aligned}$$

Similarly, with high probability no less than $1 - cn^{-3}\log n$, we have

$$\|V_\xi\|_{2,\infty} \le (\mu + C)\sqrt{\frac{r}{n}}.$$

Therefore, with high probability no less than $1 - cN^{-3}\log N$, where $N = \min\{m, n\}$, ξ is incoherent with incoherence parameter bounded by $C+\mu$. □

Theorem 6.3.1 can be interpreted as follows. Intuitively, when L_* is incoherent, any tangent element $\xi \in T_{L_*}$ is at least somewhat incoherent, because ξ can be written as $\xi = UA^\top + BV^\top$ for some $A \in \mathbb{R}^{n\ \times r}$ and $B \in \mathbb{R}^{m\times r}$ and it seems that $UA^\top$ and $BV^\top$ are one-sided incoherent. Theorem 6.3.1 tells us that a one-sided incoherent matrix is with high probability truly incoherent.

The next theorem is a deterministic result on the incoherence in the tangent space, expressed in terms of the tangent space projection.

Theorem 6.3.2. *Let P_{T^*} be the shorthand for the tangent space projection $P_{T_{L_*}}$ at the ground truth L_*, and let P_{Ω^*} be the projection onto the support of ground truth S_*. Under Assumptions 6.1.1, we have*

$$\|P_{T^*}P_{\Omega^*}P_{T^*}\|_2 \le \sqrt{2p\mu^2 r}\|P_{T^*}\|_2.$$

Proof. To prove the theorem is equivalent to proving that for any ξ in the tangent space T_{L_*},

$$\|P_{\Omega^*}(\xi)\|_F^2 \le 2p\mu^2 r\|\xi\|_F^2.$$

Note that each ξ can be expressed as $\xi = UA^\top + BV^\top$ with A and B satisfying

$$\|UA^\top + BV^\top\|_F^2 = \|UA^\top\|_F^2 + \|BV^\top\|_F^2.$$

To see this, consider the decomposition

$$\xi = \begin{pmatrix} U, & \widetilde{U} \end{pmatrix} \begin{pmatrix} M & N_1 \\ N_2 & 0 \end{pmatrix} \begin{pmatrix} V^\top \\ \widetilde{V}^\top \end{pmatrix},$$

and let $A = (M,\ N_1) \begin{pmatrix} V^\top \\ \widetilde{V}^\top \end{pmatrix}$ and $B = (U,\ \widetilde{U})N_2$. Thus, it suffices to prove that

$$\|P_{\Omega^*}(\xi)\|_F^2 \le 2p\mu^2 r(\|UA^\top\|_F^2 + \|BV^\top\|_F^2).$$

We first compare $\|UA^\top\|_F^2$ and $\|P_{\Omega^*}(UA^\top)\|_F^2$. Write

$$A = \begin{bmatrix} a_1^\top \\ \vdots \\ a_n^\top \end{bmatrix}.$$

Then

$$\|UA^\top\|_F^2 = \sum_{i=1}^n \|Ua_i\|_2^2 = \sum_{i=1}^n \|a_i\|_2^2.$$

On the other hand,

$$\|P_{\Omega^*}(UA^\top)\|_F^2 = \sum_{i=1}^n \|P_{\Omega(:,i)}(Ua_i)\|_2^2.$$

Since

$$\|Ua_i\|_\infty \le \|U\|_{2,\infty}\|a_i\|_2 \le \mu\sqrt{\frac{r}{m}}\|a_i\|_2,$$

we have

$$\|P_{\Omega(:,i)}(Ua_i)\|_2^2 \le (pm)\left(\mu\sqrt{\frac{r}{m}}\|a_i\|_2\right)^2 = p\mu^2 r\|a_i\|_2^2.$$

Putting all columns together, we get

$$\|P_{\Omega^*}(UA^\top)\|_F^2 \le p\mu^2 r\|UA^\top\|_F^2.$$

Similarly,

$$\|P_{\Omega^*}(BV^\top)\|_F^2 \le p\mu^2 r\|BV^\top\|_F^2.$$

Thus,

$$\begin{aligned}\|P_{\Omega^*}(\xi)\|_F^2 &= \|P_{\Omega^*}(UA^\top + BV^\top)\|_F^2 \\ &\le 2\left(\|P_{\Omega^*}(UA^\top)\|_F^2 + \|P_{\Omega^*}(BV^\top)\|_F^2\right) \\ &\le 2p\mu^2 r\left(\|UA^\top\|_F^2 + \|BV^\top\|_F^2\right) \\ &\le 2p\mu^2 r\|UA^\top + BV^\top\|_F^2.\end{aligned}$$

This gives us the desired result. □

We emphasize that Theorem 6.3.2 is a deterministic result. There is no lower bound for $P_{T^*}P_{\Omega^*}P_{T^*}$, and counterexamples are easy to construct, so it is not a concentration result. We include this result here because Assumptions 6.1.1 are also deterministic without any underlying probabilistic constructions, and a corresponding incoherence theorem might be of theoretical interest.

6.4 Local optimality

In this section, we present the first main result of the chapter, which gives that L_* is a local minimizer of the function $f(L)$ on the low-rank matrix manifold $\mathcal{M}_r$.

The main theorem of this section is as follows.

Theorem 6.4.1 (Local optimality)**.** *For $r = 1$, under Assumptions 6.1.1, L_* is a local minimizer of $f(L)$ on $\mathcal{M}_1$ in the sense that*

1) *For all L such that $\|L - L_*\|_{l_1} \le \frac{\sigma_r^2}{120r\sigma_1\sqrt{mn}}$, we have that $f(L) - f(L_*) > 0$;*

2) *Furthermore, if Condition 6.1.3 is satisfied, then for all L such that $\|L - L_*\|_{l_1} \le \frac{c^2\sigma_r^2\sqrt{mn}}{120r\sigma_1}$ and $\|L - L_*\|_2 \le \frac{\sigma_r}{2}$, we have that $f(L) - f(L_*) > 0$.*

Here σ is the singular value of L_* and C is an independent constant.

We remark that while Theorem 6.4.1 only applies to $r = 1$, in numerical experiments, the local optimality seems to hold true for $r > 1$ as well. Establishing this result for general r is left for future work. Also, the dependence on m and n could be improved if expressed in other norms, but we use the l_1 norm because this is the measure of distance for convergence in our Riemannian subgradient algorithm.

The proof of Theorem 6.4.1 relies on the following lemmas, and is deferred to the end of this subsection.

Lemma 6.4.2. *A convex function on a bounded convex polytope C achieves its maximum at the extreme points of C.*

The proof is obvious by definition of convex functions and convex polytopes.

Lemma 6.4.3 (adapted from [61, Theorem 3.14])**.** *Let $C \subset \mathbb{R}^n$ be a closed convex polytope. Then there exists $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $A' \in \mathbb{R}^{\times n}$, $b' \in \mathbb{R}^{\times n}$, such that*

$$C = \{x \in \mathbb{R}^n : \; Ax + b \le 0, \quad A'x + b' = 0\},$$

where the inequality holds in a componentwise sense. Moreover, every face[1] *$\mathcal{F}$ of the convex polytope C can be represented by forcing some inequality constraints of C into equality constraints, i.e., there exist $A_1 \in \mathbb{R}^{k\times n}$, $A_2 \in \mathbb{R}^{(m-k)\times n}$ and $b_1 \in \mathbb{R}^k$, $b_2 \in \mathbb{R}^{m-k}$, such that*

$$\mathcal{F} = \{x \in \mathbb{R}^n : \; A_1x + b_1 \le 0, \; A_2x + b_2 = 0, \; A'x + b' = 0\},$$

where $\begin{bmatrix} A_1 & b_1 \\ A_2 & b_2 \end{bmatrix} = \Pi \begin{bmatrix} A & b \end{bmatrix}$ and Π is a permutation matrix.

Using the lemma, we can determine the representation of the vertices of a convex polytope by making as many inequality constraints into equality constraints as possible. We will show the local optimality of L_* by finding the vertices of the polytope $C_1 := \{\xi : \xi \in T_{L_*}, \|\xi\|_{l_1} \le 1\}$ and finding an upper bound for $\frac{\|P_{\Omega^*}(\xi)\|_{l_1}}{\|\xi\|_{l_1}}$ there. More specifically, $\frac{\|P_{\Omega^*}(\xi)\|_{l_1}}{\|\xi\|_{l_1}} < \frac{1}{3}$.

Lemma 6.4.4. *The set $C_1 = \{\xi : \xi \in T_{L_*}, \|\xi\|_{l_1} \le 1\}$ is a convex polytope.*

Proof. The inequality constraint $\|\xi\|_{l_1} \le 1$ is equivalent to the following constraints:

$$\sum_{i,j} \epsilon_{i,j}\xi_{i,j} \le 1, \quad \forall \epsilon = (\epsilon_{i,j})_{m\times n} \in \{\pm 1\}^{m\times n}.$$

[1] Vertices are seen as zero-dimensional faces.

Thus C_1 is a subset of $\mathbb{R}^{m\times n}$ described by a set of equality and inequality constraints. So C_1 is a convex polyhedron. Moreover, it is a bounded polyhedron due to the constraint $\|\xi\|_{l_1} \leq 1$. Therefore, C_1 is a convex polytope. □

Denote $C_0 := \{\xi : \xi \in T_{L_*}, \|\xi\|_{l_1} = 1\}$, where $\text{rank}(L_*) = 1$. One can see that C_0 is the boundary of C_1, and is itself a collection of convex polytopes. Moreover, the vertices of C_0 are the same as those of C_1.

Proof of Lemma 6.4.6

The following two lemmas describe the structure of a vertex ξ and the behavior of $P_{\Omega^*}(\xi)$ there. We first prove the following intermediate lemma about a property of the zero patterns of the vertices of the polytope C_1.

Lemma 6.4.5. *Let ξ be a vertex of the polytope C_1, and let $\Omega^c(\xi)$ denote the index set of the zero entries in ξ. Then ξ is a matrix that contains as many zeros as possible, in the sense that there does not exist another vertex ξ' such that $\Omega^c(\xi) \subsetneq \Omega^c(\xi')$.*

Proof. By Lemma 6.4.3, the faces of C_1 are determined by forcing some inequality constraints of C_1 into equality constraints. The vertices are the zero-dimensional faces of C_1, and they can be determined through the same procedure. We now study the equality constraints created in this way. It is easy to see that there must be at least two such equalities. Let $k = (i, j)$ denote the indices. Let K denote the index set of ϵ that flip signs between these two equalities. Then

$$\begin{aligned} \sum_{k\in K} \epsilon_k \xi_k + \sum_{k\in K^c} \epsilon_k \xi_k &= 1, \\ \sum_{k\in K} (-\epsilon_k) \xi_k + \sum_{k\in K^c} \epsilon_k \xi_k &= 1. \end{aligned} \tag{6.4}$$

Therefore,

$$\sum_{k\in K} \epsilon_k \xi_k = 0, \qquad \sum_{k\in K^c} \epsilon_k \xi_k = 1.$$

We now look at all the inequality constraints with the same choices of signs in K^c, but different choice of signs in K. They are

$$\sum_{k\in K} \epsilon'_k \xi_k + \sum_{k\in K^c} \epsilon_k \xi_k \leq 1, \quad \forall (\epsilon'_k)_{k\in K} \in \{\pm 1\}^{|K|}.$$

This implies

$$\sum_{k\in K} \epsilon'_k \xi_k \leq 0, \quad \forall (\epsilon'_k)_{k\in K} \in \{\pm 1\}^{|K|}.$$

Since the signs $(\epsilon'_k)_{k\in I}$ are arbitrary, we let $\epsilon'_k = \text{sign}(\xi_k)$. This gives

$$\sum_{k\in K} |\xi_k| \leq 0,$$

which implies

$$\xi_k = 0, \quad \forall k \in K.$$

In other words, the effect of forcing inequalities constraints into equalities is inducing zeros in ξ.

On the other hand, one can show that all the zeros in ξ are induced by equality constraints. To see this, let J denote the index set of zeros that are not in the previous index set K. In other words, $J = \Omega^c(\xi)\backslash K$. Rewrite the previous pair of equalities (6.4) as

$$\begin{aligned} \sum_{k\in K} \epsilon_k \xi_k + \sum_{k\in J} \epsilon_k \xi_k + \sum_{k\in \Omega(L_*)} \epsilon_k \xi_k = 1, \\ \sum_{k\in K} (-\epsilon_k) \xi_k + \sum_{k\in J} \epsilon_k \xi_k + \sum_{k\in \Omega(L_*)} \epsilon_k \xi_k = 1. \end{aligned} \tag{6.5}$$

Since $\xi_k = 0 \; \forall k \in J$, we immediately have

$$\sum_{k\in K} (-\epsilon_k) \xi_k + \sum_{k\in J} (-\epsilon_k) \xi_k + \sum_{k\in \Omega(L_*)} \epsilon_k \xi_k = 1. \tag{6.6}$$

Therefore, Equation (6.6) and the first equation of (6.5) make a pair, and $K \cup J$ can be interpreted as zeros induced by a pair of equality constraints. To conclude, we have shown that all the zero entries in ξ are induced by pairs of equality constraints as above.

Moreover, it can be shown that more zeros always imply a lower-dimensional face. In fact, let the index set of the zeros be $\Omega^*(L_*)$, then the equality constraints are exactly the following:

$$\sum_{k\in \Omega^*(L_*)} \epsilon_k \xi_k + \sum_{k\in \Omega(L_*)} \text{sign}(\xi_k) \xi_k = 1.$$

The other inequality constraints are strict inequalities because one cannot flip the sign before a ξ_k that is nonzero.

Now suppose we have two points in two faces

$$\xi \in \mathcal{F}, \quad \xi' \in \mathcal{F}',$$

and suppose the zero pattern of ξ is a subset of the zero pattern of ξ', i.e.,

$$\Omega^c(\xi) \subsetneq \Omega^c(\xi'),$$

then we must have $\mathcal{F}' \subsetneq \mathcal{F}$, i.e., $\mathcal{F}'$ is a sub-face of the face $\mathcal{F}$. This means that $\mathcal{F}$ cannot be a zero-dimensional face, so ξ cannot be a vertex. □

We are now ready to show that the vertices actually have a block form under permutations.

Lemma 6.4.6. *Let $\mathcal{V}$ be the set of vertices of the convex polytope C_1 (equivalently, C_0). Then every vertex in $\mathcal{V}$ satisfies a certain block form up to permutations. Specifically,*

$$\mathcal{V} \subset \left\{ \xi : \xi = \Pi_1 \begin{pmatrix} \mathbf{0} & \xi^{(1)} \\ \xi^{(2)} & \mathbf{0} \end{pmatrix} \Pi_2, \quad \textit{where } \Pi_1, \Pi_2 \textit{ are permutation matrices} \right\}.$$

Proof of Lemma 6.4.6. Let $L_* = u\sigma v^\top$ be the SVD of L_*, where $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ are vectors in the case $\operatorname{rank}(L_*) = 1$. Then every ξ in the tangent space can be expressed as $\xi = ua^\top + bv^\top$, where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. Let u_i denote the ith entry of the vector u, and similarly for other vectors. Let ξ_{ij} denote the (i, j) entry of ξ. Then $\xi_{ij} = u_i a_j + b_i v_j$.

First assume for simplicity that $u_i \neq 0$ for all $i = 1, \ldots, m$ and $v_j \neq 0$ for all $j = 1, \ldots, n$. We will prove the following claim:

$$\text{If } \xi_{ij} = \xi_{kj} = \xi_{il} = 0, \text{ then } \xi_{kl} = 0.$$

To see this, note that $\xi_{ij} = \xi_{kj} = \xi_{il} = 0$ implies

$$\begin{aligned} u_i a_j + b_i v_j &= 0, \\ u_i a_l + b_i v_l &= 0, \\ u_k a_j + b_k v_j &= 0. \end{aligned}$$

Since u_i, u_k, v_j and v_l are nonzero, either $a_j = b_i = a_l = b_k = 0$, or $\frac{a_j}{a_l} = \frac{v_j}{v_l}$. In either case, $\xi_{kl} = u_k a_l + b_k v_l = 0$.

Essentially, the above claim guarantees that the zeros in ξ always come in blocks under permutation.

Next, we show that for any ξ, there exists another ξ' in the form of $\xi' = \Pi_1 \begin{pmatrix} \mathbf{0} & \xi'^{(1)} \\ \xi'^{(2)} & \mathbf{0} \end{pmatrix} \Pi_2$, such that the zero set of ξ is a subset of the zero set of ξ', i.e., $\Omega^c(\xi) \subset \Omega^c(\xi')$.

To see this, we show that we can construct $\xi' = \Pi_1 \begin{pmatrix} \mathbf{0} & \xi'^{(1)} \\ \xi'^{(2)} & \mathbf{0} \end{pmatrix} \Pi_2$ for any block division. In fact, denote the indices of the upper-left block after permutation as $I \times J$, and let

$$\tilde{u} = \Pi_1^{-1} u, \quad \tilde{v} = \Pi_2^{-\top} v, \quad \tilde{a} = \begin{pmatrix} \tilde{v}_J \\ 0 \end{pmatrix} \Pi_2, \quad \tilde{b} = \Pi_1 \begin{pmatrix} -\tilde{u}_I \\ 0 \end{pmatrix},$$

then

$$\tilde{\xi} = u\tilde{a}^\top + \tilde{b}v^\top = \Pi_1 \left(\tilde{u} \begin{pmatrix} \tilde{v}_J^\top, & 0 \end{pmatrix} + \begin{pmatrix} -\tilde{u}_I \\ 0 \end{pmatrix} \tilde{v}^\top \right) \Pi_2 = \Pi_1 \begin{pmatrix} 0 & -\tilde{u}_I \tilde{v}_{J^c}^\top \\ \tilde{u}_{I^c} \tilde{v}_J^\top & 0 \end{pmatrix} \Pi_2.$$

Take the normalization $\xi' = \tilde{\xi}/\|\tilde{\xi}\|_{l_1}$ and ξ' is an element in C_0 that satisfies the block form after permutation.

By Lemma 6.4.3, the vertices of C_1 are found by forcing as many of the inequality constraints in C_1 into equality constraints as possible, until there is a unique solution. By Lemma 6.4.2, this translates to forcing there to be as many zeros as possible. Since for ξ that are not in the block form, there always exists ξ' that has more zeros and are in the block form, we conclude that any vertex must have the block form.

In the case where u or v has zero entries, simply let the corresponding entries in b or a be zero. For example, if $u_i = 0$, we let $b_i = 0$ so that $\xi_{ij} = 0$ for any j. This construction can max out the zeros in ξ. It is easy to see that the same block form still applies, and the conclusion in Lemma 6.4.6 holds true. □

Proof of Lemma 6.4.7

Lemma 6.4.7. *For $r = 1$, under Assumptions 6.1.1, we have that for all $\xi \in T_{L_*}$,*

$$\frac{\|P_{\Omega^*}(\xi)\|_{l_1}}{\|\xi\|_{l_1}} \leq \frac{1}{3}.$$

The proof of this lemma is very technical. To build up some intuition, we can look at the cases when ξ only has the $UA^\top$ part or the $BV^\top$ part.

Example 6.4.8. If $\xi = UA^\top$, since U is incoherent, $\xi = UA^\top$ is row-wise spread-out. As the sparse support set Ω^* is row-wise sparse, it is intuitively correct that $\|P_{\Omega^*}(\xi)\|_{l_1}$ only takes up a small portion of $\|\xi\|_{l_1}$.

Example 6.4.9. If $\xi = BV^\top$, since V is incoherent, $\xi = BV^\top$ is column-wise spread-out. As the sparse support set Ω^* is column-wise sparse, it is also intuitively correct that $\|P_{\Omega^*}(\xi)\|_{l_1}$ only takes up a small portion of $\|\xi\|_{l_1}$.

When it comes to the case when $\xi = UA^\top + BV^\top$, the proof still follows the same basic idea. It makes use of the incoherent structure of U and V, and proves a delicate balancing of the entries of two nonzero blocks.

Proof of Lemma 6.4.7. To prove the lemma, it suffices to solve the following maximization problem

$$\max_{\xi \in T_{L^*},\, \|\xi\|_{l_1}=1} \|P_{\Omega^*}(\xi)\|_{l_1}$$

and show that the maximum value is upper bounded by $\frac{1}{3}$.

By Lemma 6.4.2, a convex function on a convex set takes its maximum at the extreme points of the convex set. Since $\|P_{\Omega^*}(\xi)\|_{l_1}$ is a convex function in ξ, and the constraint set is C_0 which is a collection of convex polytopes with vertex set $\mathcal{V}$, we only need to evaluate $\|P_{\Omega^*}(\xi)\|_{l_1}$ at $\xi \in \mathcal{V}$.

From Lemma 6.4.6, we know that every vertex ξ in the vertex set $\mathcal{V}$ must take the form $\Pi_1 \begin{pmatrix} \mathbf{0} & \xi^{(1)} \\ \xi^{(2)} & \mathbf{0} \end{pmatrix} \Pi_2$, where Π_1, Π_2 are permutation matrices. In the proof of Lemma 6.4.6, we have constructed some ξ' that takes such block form, where each ξ' is a normalization of the following $\tilde{\xi}$:

$$\tilde{\xi} = u\tilde{a}^\top + \tilde{b}v^\top = \Pi_1 \left(\tilde{u} \begin{pmatrix} \tilde{v}_J^\top, & 0 \end{pmatrix} + \begin{pmatrix} -\tilde{u}_I \\ 0 \end{pmatrix} \tilde{v}^\top \right) \Pi_2 = \Pi_1 \begin{pmatrix} 0 & -\tilde{u}_I \tilde{v}_{J^c}^\top \\ \tilde{u}_{I^c} \tilde{v}_J^\top & 0 \end{pmatrix} \Pi_2.$$

We now prove that in fact every $\tilde{\xi}$ that takes this block form must follow this construction. For simplicity assume that $\Pi_1 = \Pi_2 = \mathrm{id}$, then $\tilde{u} = u$ and $\tilde{v} = v$. Suppose that $\tilde{\xi} = ua^\top + bv^\top = \begin{pmatrix} \mathbf{0} & \xi^{(1)} \\ \xi^{(2)} & \mathbf{0} \end{pmatrix}$. Denote the indices of the upper left block after permutation as $I \times J$. Then

$$u = \begin{pmatrix} u_I \\ u_{I^c} \end{pmatrix}, \quad v = \begin{pmatrix} v_J \\ v_{J^c} \end{pmatrix}, \quad a = \begin{pmatrix} a_J \\ a_{J^c} \end{pmatrix}, \quad b = \begin{pmatrix} b_I \\ b_{I^c} \end{pmatrix},$$
$$u_I a_J^\top + b_I v_J^\top = 0, \quad u_{I^c} a_{J^c}^\top + b_{I^c} v_{J^c}^\top = 0.$$

More specifically, we have

$$u_i a_j + b_i v_j = 0, \quad \forall (i, j) \in (I \times J) \cup (I^c \times J^c).$$

Suppose that all entries of u and v are nonzero. If $|I| \geq 2$ and $|J| \geq 2$, it can be deduced that $\frac{a_j}{a_l} = \frac{v_j}{v_l}$ for any $j,\ l \in J$, and $\frac{b_i}{b_k} = \frac{u_i}{u_k}$ for any $i,\ k \in I$. In other words, a_J is just a rescaling of v_J and b_I is just a rescaling of u_I. Moreover, b_I is determined by a_J because they are linked by the relation $u_i a_j + b_i v_j = 0$. We can thus let $a_J = C_1 \cdot v_J$, and we have $b_I = -C_1 \cdot u_I$. The same thing applies to the $I^c \times J^c$ block, and we have $a_{J^c} = C_2 \cdot v_{J^c}$ and $b_{J^c} = -C_2 \cdot u_{I^c}$. Thus,

$$\tilde{\xi} = ua^\top + bv^\top = \begin{pmatrix} 0 & (C_2 - C_1) u_I v_{J^c}^\top \\ (C_1 - C_2) u_{I^c} v_J^\top & 0 \end{pmatrix}.$$

The constant $C_2 - C_1$ can be determined by the normalization constraint $\|\xi\|_{l_1} = 1$. This means that the ξ that takes this block form is unique. Since we have already found an explicit form in the proof of Lemma 6.4.6, this is the unique solution.

When $|I| = 1$, we still have $\frac{a_j}{a_l} = \frac{v_j}{v_l}$ for any $j,\ l \in J$, and b_I is still a rescaling of u_I because both are scalars. When an entry of u or v is zero, from the proof of Lemma 6.4.6, we let the corresponding entry of b or a be zero, and the rescaling relation still holds. Thus, the previous conclusion is still valid.

We now look at the block form that we have found, and study the value of $\|P_{\Omega^*}(\xi)\|_{l_1}$ where Ω^* is the support of L_*. For simplicity we still assume $\Pi_1 = \Pi_2 = \mathrm{id}$, and rewrite a vertex candidate ξ as

$$\xi = C \cdot \begin{pmatrix} 0 & u_I v_{J^c}^\top \\ u_{I^c} v_J^\top & 0 \end{pmatrix}.$$

We are interested in the following ratio

$$\frac{\|P_{\Omega^*}(\xi)\|_{l_1}}{\|\xi\|_{l_1}} = \frac{\|P_{\Omega^*}(u_I v_{J^c}^\top)\|_{l_1} + \|P_{\Omega^*}(u_{I^c} v_J^\top)\|_{l_1}}{\|u_I v_{J^c}^\top\|_{l_1} + \|u_{I^c} v_J^\top\|_{l_1}}.$$

By Assumptions 6.1.1, u and v are incoherent and Ω^* is sparse. We look at what this implies for the mass distribution of $u_I v_{J^c}^\top$, $u_{I^c} v_J^\top$, and their projection onto the support set Ω^*.

Let us look at u as an example. Since u is incoherent, $|u_i| \leq \frac{\mu}{\sqrt{m}}$ for each i. On the other hand, $\|u\|_2 = 1$, i.e., $\sum_i |u_i|^2 = 1$. Therefore, we know that there are at least $\frac{m}{\mu^2}$ nonzero entries in u, and $\|u\|_1 \geq \frac{\sqrt{m}}{\mu}$.

For any index division $I \cup I^c$, either there are many nonzeros in I, or there are many in I^c. For any $\gamma \leq \frac{1}{2}$, either $\sum_{i \in I} |u_i| \geq \gamma \frac{\sqrt{m}}{\mu}$ or $\sum_{i \in I^c} |u_i| \geq \gamma \frac{\sqrt{m}}{\mu}$ must be true. The same thing applies to v as well. In the following, we discuss the cases determined by the above criterion, and look into what that implies for $P_{\Omega^*}(\xi)$ and ξ. We take $\gamma = \frac{1}{3}$.

(1) If $\sum_{i \in I} |u_i| \geq \gamma \frac{\sqrt{m}}{\mu}$ and $\sum_{j \in J} |v_i| \geq \gamma \frac{\sqrt{n}}{\mu}$:

In this case, u has large mass in I and v has large mass in J. Consider the upper-right block $u_I v_{J^c}^\top$ in ξ. For the jth column in this block,

$$\|\xi_{I,j}\|_{l_1} = \sum_{i \in I} |u_i v_j| \geq \gamma \frac{\sqrt{m}}{\mu} |v_j|.$$

On the other hand, since Ω^* is sparse, Ω^* takes up at most pm entries in this column. In each entry, $|u_i|$ is upper bounded by $\frac{\mu}{\sqrt{m}}$ due to incoherence. Thus

$$\|P_{\Omega^*}(\xi_{I,j})\|_{l_1} = \sum_{i \in I:\, (i,j) \in \Omega^*} |u_i v_j| \leq pm \frac{\mu}{\sqrt{m}} |v_j| = p\mu\sqrt{m}|v_j|.$$

Therefore,

$$\frac{\|P_{\Omega^*}(\xi_{I,j})\|_{l_1}}{\|\xi_{I,j}\|_{l_1}} \leq \frac{p\mu\sqrt{m}|v_j|}{\gamma \frac{\sqrt{m}}{\mu}|v_j|} = \frac{p\mu^2}{\gamma}.$$

Since this applies to every column, for the whole upper-right block, we have

$$\frac{\|P_{\Omega^*}(u_I v_{J^c}^\top)\|_{l_1}}{\|u_I v_{J^c}^\top\|_{l_1}} \leq \frac{p\mu^2}{\gamma}.$$

For the lower-left block, we can derive the same bound by looking at each row and using the fact that v has large mass in J. This gives

$$\frac{\|P_{\Omega^*}(u_{I^c} v_J^\top)\|_{l_1}}{\|u_{I^c} v_J^\top\|_{l_1}} \leq \frac{p\mu^2}{\gamma}.$$

Therefore,

$$\frac{\|P_{\Omega^*}(\xi)\|_{l_1}}{\|\xi\|_{l_1}} \le \frac{p\mu^2}{\gamma} \le \frac{1}{12} < \frac{1}{3}.$$

(2) If $\sum_{i\in I}|u_i| \ge \gamma\frac{\sqrt{m}}{\mu}$ and $\sum_{i\in I^c}|u_i| \ge \gamma\frac{\sqrt{m}}{\mu}$, while $\sum_{j\in J}|v_i| < \gamma\frac{\sqrt{n}}{\mu}$:

In this case, we must have $\sum_{j\in J^c}|v_i| \ge \gamma\frac{\sqrt{n}}{\mu}$. Therefore this case can be proved in the same way as case (1), except that for the upper-right block we calculate by row and for the lower-left block we calculate by column.

(3) If $\sum_{i\in I^c}|u_i| < \gamma\frac{\sqrt{m}}{\mu}$, and $\sum_{j\in J}|v_i| < \gamma\frac{\sqrt{n}}{\mu}$:

In this case, u has very large mass in I and v has very large mass in J^c. This means that the upper-right block $u_I v_{J^c}^\top$ has very large mass and dominates over the lower-left block $u_{I^c} v_J^\top$. For the upper-right block, we use the same technique as case (1) to derive that

$$\frac{\|P_{\Omega^*}(u_I v_{J^c}^\top)\|_{l_1}}{\|u_I v_{J^c}^\top\|_{l_1}} \le \frac{p\mu^2}{\gamma}.$$

For the lower-left block, we now show that its mass is so small that it does not tweak the ratio too much. More specifically,

$$\|u_{I^c} v_J^\top\|_{l_1} = \left(\sum_{I_c}|u_i|\right)\left(\sum_{J}|v_j|\right) \le \gamma\frac{\sqrt{m}}{\mu}\cdot\gamma\frac{\sqrt{n}}{\mu} = \gamma^2\frac{\sqrt{mn}}{\mu^2}.$$

On the other hand,

$$\|u_I v_{J^c}^\top\|_{l_1} = \left(\sum_{I}|u_i|\right)\left(\sum_{J^c}|v_j|\right) \ge (1-\gamma)\frac{\sqrt{m}}{\mu}\cdot(1-\gamma)\frac{\sqrt{n}}{\mu} = (1-\gamma)^2\frac{\sqrt{mn}}{\mu^2}.$$

Therefore, we have

$$\begin{aligned}\frac{\|P_{\Omega^*}(\xi)\|_{l_1}}{\|\xi\|_{l_1}} &= \frac{\|P_{\Omega^*}(u_I v_{J^c}^\top)\|_{l_1} + \|P_{\Omega^*}(u_{I^c} v_J^\top)\|_{l_1}}{\|u_I v_{J^c}^\top\|_{l_1} + \|u_{I^c} v_J^\top\|_{l_1}} \le \frac{\|P_{\Omega^*}(u_I v_{J^c}^\top)\|_{l_1} + \|u_{I^c} v_J^\top\|_{l_1}}{\|u_I v_{J^c}^\top\|_{l_1}}\\ &\le \frac{p\mu^2}{\gamma} + \frac{\gamma^2}{(1-\gamma)^2} \le \frac{1}{12} + \frac{1/9}{4/9} = \frac{1}{3}.\end{aligned}$$

(4) If $\sum_{i\in I}|u_i| < \gamma\frac{\sqrt{m}}{\mu}$, while $\sum_{j\in J}|v_i| \ge \gamma\frac{\sqrt{n}}{\mu}$ and $\sum_{j\in J^c}|v_i| \ge \gamma\frac{\sqrt{n}}{\mu}$:

This case is the same as case (2).

(5) If $\sum_{i\in I}|u_i| < \gamma\frac{\sqrt{m}}{\mu}$, and $\sum_{j\in J^c}|v_i| < \gamma\frac{\sqrt{n}}{\mu}$:

This case is the same as case (3), except that now the lower-left block $u_{I^c} v_J^\top$ has very large mass and the upper-right block has small mass.

(6) If $\sum_{i\in I} |u_i| < \gamma\frac{\sqrt{m}}{\mu}$ and $\sum_{j\in J} |v_i| < \gamma\frac{\sqrt{n}}{\mu}$:

This case is the same as case (2).

To conclude, for all $\xi \in T_{L_*}$, we have $\frac{\|P_\Omega(\xi)\|_{l_1}}{\|\xi\|_{l_1}} \leq \frac{1}{3}$. □

Proof of Theorem 6.4.1

We are now ready to prove the local optimality of L_* on the manifold $\mathcal{M}_r$.

Proof of Theorem 6.4.1. For any rank-r matrix L,

$$\begin{aligned} f(L) - f(L_*) &= \|M - L\|_{l_1} - \|M - L_*\|_{l_1} \\ &= \|S_* + L_* - L\|_{l_1} - \|S_*\|_{l_1}. \end{aligned}$$

Recall that the support of S_* is Ω^*, which is p-sparse. Separating the entries in Ω^* and $(\Omega^*)^c$, we have

$$\begin{aligned} &\|S_* + L_* - L\|_{l_1} - \|S_*\|_{l_1} \\ =&\|P_{\Omega^*}(S_* + L_* - L)\|_{l_1} + \|P_{(\Omega^*)^c}(S_* + L_* - L)\|_{l_1} - \|S_*\|_{l_1} \\ \geq& - \|P_{\Omega^*}((S_* + L_* - L) - S_*)\|_{l_1} + \|P_{(\Omega^*)^c}(S_* + L_* - L)\|_{l_1} \\ =& - \|P_{\Omega^*}(L_* - L)\|_{l_1} + \|P_{(\Omega^*)^c}(L_* - L)\|_{l_1}. \end{aligned}$$

It remains the show that $\|P_{(\Omega^*)^c}(L_* - L)\|_{l_1}$ is consistently larger than $\|P_{\Omega^*}(L_* - L)\|_{l_1}$. This is intuitively correct because Ω^* is much sparser than $(\Omega^*)^c$ when p is small. To show this, let $\xi = P_{T_{L_*}}(L - L_*)$ be the tangent space projection of $L - L_*$. Then

$$\begin{aligned} \frac{\|P_{\Omega^*}(L_* - L)\|_{l_1}}{\|P_{\Omega^*}(L_* - L)\|_{l_1} + \|P_{(\Omega^*)^c}(L_* - L)\|_{l_1}} &= \frac{\|P_{\Omega^*}(L_* - L)\|_{l_1}}{\|L - L_*\|_{l_1}} \\ &\leq \frac{\|P_{\Omega^*}(\xi)\| + \|(L - L_*) - \xi\|_{l_1}}{\|\xi\|_{l_1} - \|(L - L_*) - \xi\|_{l_1}}. \end{aligned} \tag{6.7}$$

We now show $\|(L - L_*) - \xi\|_{l_1}$ is sufficiently small when L is sufficiently close to L_*. Specifically, for arbitrary L, we have

$$\begin{aligned} \|(L - L_*) - \xi\|_{l_1} &= \|(L - L_*) - P_{T_{L_*}}(L - L_*)\|_{l_1} \\ &= \|L - P_{T_{L_*}}(L)\|_{l_1} \\ &= \|P^\perp_{T_{L_*}}(L)\|_{l_1} \\ &\leq \sqrt{mn}\|P^\perp_{T_{L_*}}(L)\|_F \\ &\leq \sqrt{mn}\|P_{U_L} - P_{U_{L_*}}\|_2 \cdot \|P_{V_L} - P_{V_{L_*}}\|_2 \cdot \|L\|_F. \end{aligned} \tag{6.8}$$

It is known [126] that for any unitary invariant norm (including operator 2-norm and Frobenius norm), the following SVD perturbation bound holds:

$$\|P_{U_L} - P_{U_{L_*}}\|^2 + \|P_{V_L} - P_{V_{L_*}}\|^2 \le \frac{2}{\delta^2}\|L - L_*\|^2, \quad \text{where } \delta = \min\{\sigma_r(L_*), \sigma_r(L)\}. \tag{6.9}$$

Note that in both cases (with or without Condition 6.1.3), we always have $\|L - L_*\|_2 \le \frac{\sigma_r}{2}$. This is because in the first case without Condition 6.1.3, given $\|L - L_*\|_{l_1} \le \frac{\sigma_r^2}{120 r\sigma_1\sqrt{mn}}$, we have

$$\|L - L_*\|_2 \le \|L - L_*\|_F \le \|L - L_*\|_{l_1} \le \frac{\sigma_r}{2} \cdot \frac{\sigma_r}{60 r\sigma_1\sqrt{mn}} < \frac{\sigma_r}{2}.$$

Since $\|L - L_*\|_2 \le \frac{\sigma_r}{2}$, we have $\delta = \min\{\sigma_r(L_*), \sigma_r(L)\} \ge \frac{\sigma_r(L_*)}{2}$. Thus

$$\|P_{U_L} - P_{U_{L_*}}\|^2 + \|P_{V_L} - P_{V_{L_*}}\|^2 \le \frac{8}{\sigma_r^2}\|L - L_*\|^2.$$

Using the Frobenius norm and plugging the above into (6.7), we have

$$\|(L - L_*) - \xi\|_{l_1} \le \frac{8\sqrt{mn}}{\sigma_r^2}\|L - L_*\|_F^2\|L\|_F. \tag{6.10}$$

Since $\|L\|_F \le \|L_*\|_F + r\frac{\sigma_r}{2} \le \frac{3}{2}r\sigma_1$, and $\|L - L_*\|_F \le \|L - L_*\|_{l_1}$, we have

$$\|(L - L_*) - \xi\|_{l_1} \le \frac{12 r\sigma_1\sqrt{mn}}{\sigma_r^2}\|L - L_*\|_{l_1}^2. \tag{6.11}$$

Thus, in the first case, if $\|L - L_*\|_{l_1} < \frac{\sigma_r^2}{120 r\sigma_1\sqrt{mn}}$, we then have

$$\begin{aligned}\|(L - L_*) - \xi\|_{l_1} &< \frac{12 r\sigma_1\sqrt{mn}}{\sigma_r^2} \cdot \frac{\sigma_r^2}{120 r\sigma_1\sqrt{mn}}\|L - L_*\|_{l_1} \\ &\le \frac{1}{10}\|L - L_*\|_{l_1} \\ &\le \frac{1}{10}\left(\|(L - L_*) - \xi\|_{l_1} + \|\xi\|_{l_1}\right).\end{aligned}$$

This implies

$$\|(L - L_*) - \xi\|_{l_1} < \frac{1}{9}\|\xi\|_{l_1}.$$

By Lemma 6.4.7, $\frac{\|P_\Omega(\xi)\|_{l_1}}{\|\xi\|_{l_1}} \le \frac{1}{3}$. Plugging both of them into (6.7), we have

$$\frac{\|P_{\Omega^*}(L_* - L)\|_{l_1}}{\|P_{\Omega^*}(L_* - L)\|_{l_1} + \|P_{(\Omega^*)^c}(L_* - L)\|_{l_1}} < \frac{1/3 + 1/9}{1 - 1/9} = \frac{1}{2}.$$

Thus, $\|P_{\Omega^*}(L_* - L)\|_{l_1} < \|P_{(\Omega^*)^c}(L_* - L)\|_{l_1}$, which implies $f(L) - f(L_*) > 0$. When Condition 6.1.3 holds, i.e., when $\eta(L - L_*) = \frac{\|L-L_*\|_{l_1}}{\|L-L_*\|_F} \geq c\sqrt{mn}$, we can deduce from (6.10) that

$$\|(L - L_*) - \xi\|_{l_1} \leq \frac{12r\sigma_1\sqrt{mn}}{\sigma_r^2}\frac{1}{c^2mn}\|L - L_*\|_{l_1}^2 = \frac{12r\sigma_1}{c^2\sigma_r^2\sqrt{mn}}\|L - L_*\|_{l_1}^2.$$

In this case, we only need $\|L - L_*\|_{l_1} < \frac{c^2\sigma_r^2\sqrt{mn}}{120r\sigma_1}$ to obtain the same result. This proves the theorem. □

6.5 Linear convergence of the subgradient method

Having established the local optimality of L_* on the manifold $\mathcal{M}_r$, in this section, we establish the linear convergence rate of the Riemannian subgradient algorithm (6.2) toward L_*. The main theorem is as follows. Note that this result holds for all r and is not restricted to $r = 1$.

Theorem 6.5.1 (Linear convergence). *Let L_0 be the spectral initialization generated by Algorithm 2. Let $\{L_k\}_{k=0}$ be the sequence generated by the Riemannian subgradient algorithm* (6.2) *with stepsize $\alpha_k = \alpha_0 \cdot \beta^k$, where α_0 and $\beta < 1$ are small enough positive constants. Assume that L_* is the local minima of* (6.1). *Let Condition 6.1.3 hold for all $k \geq 0$. Assume further that*

$$p \leq \frac{c^2\sigma_r^2}{1344\mu r^3\sigma_1^2}.$$

Then $\{L_k\}_{k=0}$ converges linearly to L_*, i.e., there exists $C, c > 0$ such that $\|L_k - L_*\|_{l_1} \leq Ce^{-ck}$ for large enough k.

The proof of Theorem 6.5.1 is deferred to the end of this section. The major tool of the proof is the following lemma which establishes that sharpness and weak convexity imply the linear convergence rate of subgradient methods.

Lemma 6.5.2 ([45, Theorem 5.1]). *Assume $f(x) : \mathcal{M} \to \mathbb{R}$ is L-Lipschitz with respect to a norm $\|\cdot\|$. Let $\bar{x}$ be the local minimizer of $f(x)$ in a subset of the manifold $\mathcal{B} \subseteq \mathcal{M}$. If $f(x)$ satisfies the following two conditions:*

- *(Sharpness) For any $x \in \mathcal{B}$,*

$$f(x) - f(\bar{x}) \geq \omega \cdot \|x - \bar{x}\|. \qquad \text{(S)}$$

- *(Weak Convexity) For any $x,\ y \in \mathcal{B}$,*

$$f(y) \geq f(x) + \langle \partial f(x), y - x \rangle - \frac{\rho}{2}\|y - x\|^2. \tag{C}$$

Let the stepsize in Algorithm 1 be $\alpha_k = \lambda \cdot q^k$, where $\lambda = \frac{\gamma\omega^2}{\rho L}$, $q = \sqrt{1-(1-\gamma)\tau^2}$, where $\tau = \frac{\omega}{L}$, and $\gamma \in (0,1)$ is a fixed constant. Then we have

$$\|z_k - \bar{x}\|^2 \leq \frac{\gamma^2\omega^2}{\rho^2}\left(1-(1-\gamma)\tau^2\right)^k.$$

The constants ω in (S) and ρ in (C) are called the sharpness and weak convexity constants.

The constants ω and ρ can be interpreted as follows. A larger sharpness parameter ω implies that the function is "sharper", and the convergence is faster. In the convergence rate result, we have $(1-(1-\gamma)\tau^2)$, where $\tau = \frac{\omega}{L}$. Thus a larger ω allows us to let the stepsize decay more aggressively, which gives us faster convergence rate. On the other hand, a smaller weak convexity parameter ρ implies that the function is "more convex", and the linear convergence rate is guaranteed in a larger local region. This is because the previous convergence rate only holds in a region $T\gamma := \{x \in \mathcal{B} \subseteq \mathcal{M} : \|x - \bar{x}\| \leq \gamma\frac{\omega}{\rho}\}$ for some $\gamma \in (0,1)$. A smaller ρ allows for a larger T_γ. In fact, when $\rho = 0$, f is globally convex, and we have global convergence.

We also note that there are other types of stepsize schemes which also give linear convergence, such as the Polyak step size or the constant step length [45]. The geometric step size that we use is the most popular and suffices for our purpose.

Lemma 6.5.3. *The sharpness condition* (S) *is satisfied in the local region near L_* on $\mathcal{M}_r$. More specifically,*

1) *For all $L \in \mathcal{M}_r$ such that $\|L - L_*\|_{l_1} \leq \frac{\sigma_r^2}{168 r\sigma_1\sqrt{mn}}$ and $\|L - L_*\|_2 \leq \frac{\sigma_r}{2}$, the sharpness condition is satisfied with the $\|\cdot\|_{l_1}$ norm and $\omega = \frac{1}{9}$;*

2) *Furthermore, if Condition 6.1.3 is satisfied, then for all $L \in \mathcal{M}_r$ such that $\|L - L_*\|_{l_1} \leq \frac{c^2\sigma_r^2\sqrt{mn}}{168 r\sigma_1}$ and $\|L - L_*\|_2 \leq \frac{\sigma_r}{2}$, the sharpness condition is satisfied with $\omega = \frac{1}{9}$.*

Proof. We have shown with (6.11) in the proof of Theorem 6.4.1 that when $\|L - L_*\|_2 \le \frac{\sigma_r}{2}$, we have

$$\|(L - L_*) - \xi\|_{l_1} \le \frac{12r\sigma_1\sqrt{mn}}{\sigma_r^2}\|L - L_*\|_{l_1}^2.$$

Thus, if $\|L - L_*\|_{l_1} \le \frac{\sigma_r^2}{168r\sigma_1\sqrt{mn}}$, similar to the proof of Theorem 6.4.1, we have

$$\|(L - L_*) - \xi\|_{l_1} < \frac{12r\sigma_1\sqrt{mn}}{\sigma_r^2} \cdot \frac{\sigma_r^2}{168r\sigma_1\sqrt{mn}}\|L - L_*\|_{l_1} \le \frac{1}{14}\left(\|(L - L_*) - \xi\|_{l_1} + \|\xi\|_{l_1}\right).$$

This implies

$$\|(L - L_*) - \xi\|_{l_1} \le \frac{1}{13}\|\xi\|_{l_1},$$

which further gives us

$$\frac{\|P_{\Omega^*}(L_* - L)\|_{l_1}}{\|L_* - L\|_{l_1}} = \frac{\|P_{\Omega^*}(L_* - L)\|_{l_1}}{\|P_{\Omega^*}(L_* - L)\|_{l_1} + \|P_{(\Omega^*)^c}(L_* - L)\|_{l_1}} \le \frac{1/3 + 1/13}{1 - 1/13} = \frac{4}{9}.$$

Thus,

$$f(L) - f(L_*) \ge -\|P_{\Omega^*}(L_* - L)\|_{l_1} + \|P_{(\Omega^*)^c}(L_* - L)\|_{l_1} \ge \frac{1}{9}\|L - L_*\|_{l_1}.$$

In other words, (S) is satisfied with $\omega = \frac{1}{9}$. The case with Condition 6.1.3 can be proved similarly. □

Lemma 6.5.4. *The weak convexity* (C) *is satisfied in the local region near L_* on $\mathcal{M}_r$. Specifically, for all $L \in \mathcal{M}_r$ such that $\|L - L_*\|_2 \le \frac{\sigma_r}{2}$, f(L) satisfies weak convexity (C) with constant $\rho = \frac{12r\sigma_1\sqrt{mn}}{\sigma_r^2}$. Furthermore, under Condition 6.1.3, this can be improved to $\rho = \frac{12r\sigma_1}{c^2\sigma_r^2\sqrt{mn}}$.*

Proof. Consider $X, Y \in \mathcal{M}_r$. We have

$$\begin{aligned} & f(X) - f(Y) - \langle \partial_R f(Y), X - Y\rangle \\ =& \|M - X\|_{l_1} - \|M - Y\|_{l_1} - \langle P_{T_Y}(\partial f(Y)), X - Y\rangle \\ =& \|M - X\|_{l_1} - \|M - Y\|_{l_1} - \langle \partial f(Y), X - Y\rangle + \langle P_{T_Y}^{\perp}(\partial f(Y)), X - Y\rangle. \end{aligned}$$

The first part $\|M - X\|_{l_1} - \|M - Y\|_{l_1} - \langle \partial f(Y), X - Y\rangle \ge 0$ because $\|\cdot\|_{l_1}$ is convex in the ambient space. Now consider the second part which is the term $\langle P_{T_Y}^{\perp}(\partial f(Y)), X - Y\rangle$. We have

$$\langle P_{T_Y}^{\perp}(\partial f(Y)), X - Y\rangle = \langle \partial f(Y), P_{T_Y}^{\perp}(X - Y)\rangle \ge -\|\partial f(Y)\|_{l_\infty} \cdot \|P_{T_Y}^{\perp}(X - Y)\|_{l_1}.$$

It is easy to see that

$$\|\partial f(Y)\|_{l_\infty} = \|\widetilde{\text{sign}}(Y - M)\|_{l_\infty} \leq 1.$$

Thus, we have

$$f(X) - f(Y) - \langle \partial_R f(Y), X - Y\rangle \geq -\|P_{T_Y}^{\perp}(X - Y)\|_{l_1}.$$

To establish the weak convexity condition, it now suffices to compare $\|P_{T_Y}^{\perp}(X-Y)\|_{l_1}$ with $\|X - Y\|_{l_1}^2$. Similar to the proof of (6.11), for any X, Y in the region $\{L \in \mathcal{M}_r : \|L - L_*\| \leq \frac{\sigma_r}{2}\}$,

$$\|P_{T_Y}^{\perp}(X - Y)\|_{l_1} \leq \frac{12r\sigma_1\sqrt{mn}}{\sigma_r^2}\|X - Y\|_{l_1}^2.$$

Therefore, (C) is satisfied with

$$\rho = \frac{12r\sigma_1\sqrt{mn}}{\sigma_r^2}.$$

When Condition 6.1.3, is satisfied, we have

$$\|P_{T_Y}^{\perp}(X - Y)\|_{l_1} \leq \frac{12r\sigma_1}{c^2\sigma_r^2\sqrt{mn}}\|X - Y\|_{l_1}^2.$$

This gives

$$\rho = \frac{12r\sigma_1}{c^2\sigma_r^2\sqrt{mn}},$$

which is a smaller weak convexity constant. □

Theorem 6.5.5 (Local linear convergence)**.** *In the local neighborhood $\{L \in \mathcal{M}_r : \|L - L_*\|_{l_1} \leq \frac{\sigma_r^2}{168r\sigma_1\sqrt{mn}}\}$, we have that L_k converges linearly to L_*. If Condition 6.1.3 is satisfied, the local neighborhood can be extended to $\{L \in \mathcal{M}_r : \|L - L_*\|_{l_1} \leq \frac{c^2\sigma_r^2\sqrt{mn}}{168r\sigma_1}, \|L - L_*\|_2 \leq \frac{\sigma_r}{2}\}$.*

Proof. This is a direct consequence of Lemma 6.5.3, Lemma 6.5.4, and Lemma 6.5.2. □

The question remains as to whether the initialization is in the local region required in Theorem 6.5.5. When it comes to the initialization, we first have the following lemma characterizing the distance between the initial point and the ground truth.

Lemma 6.5.6 ([131, Theorem 1])**.** *Under Assumptions 6.1.1, the initialization L_0 is close to L_* in the spectral norm. Specifically,*

$$\|L_0 - L_*\|_2 \leq 8p\mu r \cdot \sigma_1.$$

We are now ready to prove Theorem 6.5.1, which is the linear convergence of Algorithm 1 with the initialization as in Algorithm 2

Proof of Theorem 6.5.1. To prove the linear convergence of the Riemannian subgradient using the spectral initialization, it remains to show that under the assumptions of Theorem 6.5.1, the spectral initialization L_0 is in the local neighborhood of Theorem 6.5.5, and remains in there along the whole trajectory.

We first show that the spectral initialization L_0 is in the desired local neighborhood. Since $p \leq \frac{c^2\sigma_r^2}{1344\mu r^3\sigma_1^2}$, we have

$$\begin{aligned} \|L_0 - L_*\|_{l_1} &\leq \sqrt{mn}\|L_0 - L_*\|_F \leq r\sqrt{mn}\|L_0 - L_*\|_2 \\ &\leq r\sqrt{mn} \cdot 8p\mu r\sigma_1 \\ &\leq 8\mu r^2\sigma_1\sqrt{mn} \cdot \frac{c^2\sigma_r^2}{1344\mu r^3\sigma_1^2} \\ &\leq \frac{c^2\sigma_r^2\sqrt{mn}}{168r\sigma_1}. \end{aligned}$$

We also have

$$\|L_0 - L_*\|_2 \leq 8p\mu r\sigma_1 \leq 8\mu r\sigma_1 \cdot \frac{c^2\sigma_r^2}{1344\mu r^3\sigma_1^2} = \frac{c^2\sigma_r^2}{168r^2\sigma_1} < \frac{\sigma_r}{2}.$$

Thus L_0 is in the local neighborhood $\{L \in \mathcal{M}_r : \|L - L_*\|_{l_1} \leq \frac{c^2\sigma_r^2\sqrt{mn}}{168r\sigma_1}, \|L - L_*\|_2 \leq \frac{\sigma_r}{2}\}$.

Next, we prove that the whole trajectory $\{L_k\}_{k=0}$ is in this neighborhood as well. From the proof of [45, Theorem 5.1], iteratively, the sequence satisfies $\|L_k - L_*\|_{l_1}^2 \leq \frac{\gamma^2\omega^2}{\rho^2}\left(1 - (1-\gamma)\tau^2\right)^k$, which is bounded by $\frac{\gamma^2\omega^2}{\rho^2}$. In other words, if L_k, $k = 0, 1, \ldots, K$ is in the local neighborhood, then $\|L_k - L_*\|_{l_1}^2 \leq \frac{\gamma^2\omega^2}{\rho^2}$ for $0 \leq k \leq K$. Here ω and ρ are sharpness and weak convexity constants. When Condition 6.1.3 is satisfied, we have $\omega = \frac{1}{9}$ and $\rho = \frac{12r\sigma_1}{c^2\sigma_r^2\sqrt{mn}}$. Thus,

$$\frac{\gamma\omega}{\rho} = \gamma \cdot \frac{c^2\sigma_r^2\sqrt{mn}}{108r\sigma_1}.$$

With a small enough constant γ, we have $\|L_k - L_*\|_{l_1} \leq \frac{\gamma\omega}{\rho} \leq \frac{c^2\sigma_r^2\sqrt{mn}}{168r\sigma_1}$. In addition, using Condition 6.1.3, we have

$$\|L_k - L_*\|_2 \leq \|L_k - L_*\|_F \leq \frac{1}{c\sqrt{mn}}\|L_k - L_*\|_{l_1} \leq \frac{\gamma\omega}{c\rho\sqrt{mn}} = \gamma \cdot \frac{c^2\sigma_r^2}{108r\sigma_1}.$$

With a small enough γ that only depends on c, σ_1, σ_r and r, one can have $\|L_k - L_*\|_2 \leq \frac{\sigma_r}{2}$. Combined with the bound on $\|L_k - L_*\|_{l_1}$, L_k is in the local neighborhood. Using Theorem 6.5.5, we have linear convergence of Riemannian subgradient along the whole trajectory. □

6.6 Discussion

In this chapter, we have discussed the RPCA problem as an application of the low-rank matrix manifold $\mathcal{M}_r$. We solved the RPCA problem by minimizing the l_1 loss function $f(L) = \|M - L\|_{l_1}$ over $L \in \mathcal{M}_r$, and used Riemannian subgradient descent method since $f(L)$ is a nonsmooth function with well-defined subgradient. We investigated the incoherence property of the tangent space of the manifold near the ground truth L_*, and used incoherence to show that L_* is a local minimizer of $f(L)$ when the rank $r = 1$. We then looked at the convergence rate of the Riemannian subgradient algorithm, and showed that under certain assumptions, Riemannian subgradient converges to the ground truth in a linear convergence rate. In particular, under a mild condition, the spectral initialization suffices to guarantee linear convergence of the algorithm.

Our work differs from other works in the literature in that we are the first to use Riemannian subgradient for the l_1 loss function on $\mathcal{M}_r$. This demonstrates the flexibility of the low-rank matrix manifold $\mathcal{M}_r$ for different low-rank recovery problems outside of the framework in Chapter 5. Our analysis of the Conditions (S) and (C) for the linear convergence of Riemannian subgradient algorithms could potentially be extended to other problems on the manifold.

Although the local optimality result is restricted to $r = 1$, numerical evidence suggests that this is true for $r > 1$ as well. Analysis of incoherence in the tangent space is more complicated for rank $r > 1$ since the SVD factors become nontrivial. Extension of local optimality to $r > 1$ is left for future work. Another restriction on our results is Condition 6.1.3, which, though not stringent in practice, is hard to prove in theory. It would be interesting to either remove this condition or prove that it holds true with the subgradient algorithm.

Chapter 7

SOBOLEV GRADIENT DESCENT FOR A CLASS OF NONLINEAR EIGENPROBLEMS

In this chapter, we consider an application of the convergence guarantee for manifold first-order algorithms that extends beyond the low-rank matrix manifold, as it is posed on an infinite dimensional Hilbert manifold. We explore how the insights gained and tools developed for the low-rank matrix manifold $\mathcal{M}_r$ can be extended to other manifolds in broader scientific and technological fields.

Our contribution is that we establish the linear convergence rate of the Sobolev projected gradient descent (Sobolev PGD) algorithm that has been observed in numerical experiments [72]. To show this, we propose to use the Łojasiewicz inequality as a general tool for analyzing the convergence rate of gradient descent on a Hilbert manifold. Using this tool, we show that a Sobolev gradient descent method with adaptive inner product converges exponentially fast to the ground state for the Gross-Pitaevskii eigenproblem. This method can be extended to a class of general high-degree optimizations or nonlinear eigenproblems under certain conditions. We demonstrate this generalization using several examples, in particular a nonlinear Schrödinger eigenproblem with an extra high-order interaction term. Numerical experiments are presented for these problems.

Organization of the chapter. We have given a brief introduction of the problem setting in Section 1.5. The rest of the chapter is organized as follows. In Section 7.1, we discuss the background of the problem and comparison with related work in more detail. In Section 7.2, we introduce the Łojasiewicz inequality tool with mixed norms on the Hilbert manifold as an abstract convergence theorem. In Section 7.3, we establish the main result about the exponential convergence of the a_u-Sobolev gradient descent method applied to the Gross-Pitaevskii eigenproblem. Section 7.4 is devoted to the analysis of spatial discretization. In Section 7.5, we introduce several extensions of the Sobolev gradient descent to other nonlinear eigenproblems. Some numerical results are presented in Section 7.6. Finally, we make some concluding remarks in Section 7.7.

7.1 Background

The problem of interest here is the Gross-Pitaevskii eigenproblem, which seeks $\lambda \in \mathbb{R}$ and $v \in H_0^1(\Omega)$ that satisfy the following equation

$$-\Delta v + Vv + \beta|v|^2 v = \lambda v \quad \text{on } \Omega \subset \mathbb{R}^d, \tag{7.1}$$

where Ω is a bounded region in $\mathbb{R}^d$, $V(x) \geq 0$ is an external trapping potential, and $\beta \geq 0$ is a parameter describing the repulsive interaction between particles. In physics, this describes the Bose-Einstein condensate when the temperature is close to absolute zero. The eigenstate v corresponding to the smallest λ describes the ground state of this system. It has long been studied both in experiments [5] and in numerical analysis [28, 52, 73, 99].

To find the ground state v is equivalent to solving the following minimization problem:

$$\min_{\|u\|_{L^2}=1,\ u\in H_0^1(\Omega)} E(u) := \int_\Omega \left(|\nabla u|^2 + V|u|^2 + \frac{\beta}{2}|u|^4 \right) \mathrm{d}x. \tag{7.2}$$

The constraint set $\{u \in H_0^1(\Omega) : \|u\|_{L^2} = 1\}$ is the unit sphere in $H_0^1(\Omega)$. It can be seen as an infinite dimensional Hilbert manifold. Such a manifold (with additional $L^\infty(\Omega)$ constraints) will be denoted as $\mathcal{M}$ in subsequent sections.

In this chapter, we focus on a special manifold gradient descent method named the *Sobolev projected gradient descent (Sobolev PGD)*, first proposed in [72]. This method has the following iteration formula:

$$u_{n+1} = R\left((1-\tau_n)\, u_n + \tau_n \cdot \frac{(u_n, u_n)_{L^2}}{(\mathcal{G}_{u_n} u_n, \mathcal{G}_{u_n} u_n)_{a_{u_n}}} \mathcal{G}_{u_n} u_n \right). \tag{7.3}$$

A detailed discussion of Algorithm (7.3) can be found in Section 7.3.

The idea of using a discretized normalized gradient flow (DNGF) to solve Problem (7.2) can be traced back to [15]. Following this seminal work there have been a number of variants, see e.g., [42, 43, 54] and the review paper [14]. The viewpoint of (Riemannian) manifold optimization has also been explicitly adopted in [43]. Based on those methods with fixed inner products, the adaptive version of a_u-Sobolev gradient descent has recently been proposed in [72]. Despite its popularity, quantitative convergence analysis of the DNGF family has been quite lacking. The convergence rate has been either unavailable, or only proved for the gradient flow [72]. Another popular choice is the self-consistent field iteration (SCF), see e.g., [27], but rigorous global convergence rate is also difficult to establish. There are also second-order methods like the Riemannian Newton method, but they require second-order information which can be expensive to obtain.

We highlight the main differences between our work and [72]. The authors of [72] first propose the Sobolev gradient descent method (7.3). They establish the exponential convergence rate of the *time-continuous* gradient flow. But the important question of whether the *time-discrete* gradient descent also achieves optimal exponential convergence rate remains open. Our main contribution is to give a confirmatory answer to this question.

To prove the convergence rate, we introduce the Łojasiewicz inequality tool to this problem. The Łojasiewicz inequality has been discussed in Section 1.3 where it serves as a fundamental convergence tool for low-rank recovery on the low-rank matrix manifold. We rewrite it in a slightly different form in Section 7.2. Using the Łojasiewicz inequality tool, we reveal that the key to exponential convergence is the quadratic nature of the objective energy functional. In other words, regarded as a polynomial, the objective functional should behave like a degree-2 polynomial under the given manifold metric.

Although the degree of polynomial of the objective function in Problem (7.2) is formally higher than quadratic, Algorithm (7.3) changes the situation by using an adaptive inner product $a_u(\cdot,\cdot)$ instead of a fixed inner product. As a comparison, using a fixed inner product, the Łojasiewicz exponent (the θ in Theorem 7.2.1) calculated in [132] is $1/4$; while in our work, using an adaptive inner product, we have $\theta = 1/2$. The latter is more desirable according to Theorem 7.2.1. Thus, in Section 7.3, using the Łojasiewicz inequality tool, we are able to prove the exponential convergence rate of discrete time gradient descent directly.

The Łojasiewicz inequality tool also makes the Sobolev gradient descent easily applicable to general optimization of high-degree objective or eigenvalue problems other than the Gross-Pitaevskii eigenvalue problem. Its interesting property of making a high-degree polynomial behave like quadratic is not specific to a certain problem, but is general. Examples include the biharmonic Schrödinger, the nonlinear Schrödinger with a different order or extra interaction terms, and potentially some general manifold optimization problems.

In addition to the necessary regularity conditions, the only essential requirement is that the global ground state of the nonlinear problem is also the unique ground state of its *linearized* version, what we call the "double ground state" property.[1] For Problem (7.1), this property will be rigorously proved in Section 7.3. For many other problems, it is either provable, or a reasonable assumption according to numerical evidence.

Specifically, an example of nonlinear Schrödinger eigenproblem from [17] is rigorously discussed in Section 7.5. This example has an extra high-order interaction term $-\delta\Delta(|v|^2)v$ where $\delta \geq 0$. Classical methods that work for (7.1) could become inefficient or unstable for this problem. A density function reformulation $\rho := |u|^2$ was proposed in [16], but it has to treat the lack of continuity of $\nabla\sqrt{\rho}$ near 0^+ with extra regularization. Therefore the adaptive Sobolev gradient descent is advantageous for its simplicity and fast convergence.

We remark that if the domain is convex, an alternative approach to derive local linear convergence rate[2] of gradient descent methods is to use *strong convexity* (SC). This is especially popular in the finite dimensional data science problems [39]. Attempts have also been made to extend it to nonconvex settings like manifolds. Some works in this direction can be found in [1, 29]. We emphasize that our approach using the Łojasiewicz inequality has its advantages over SC, namely it applies to degenerate critical points where SC could fail, and it allows more freedom in the choice of iterative algorithms and convergence measures. A more detailed comparison of these two approaches would be of interest in future research.

7.2 Abstract convergence theorem using the Łojasiewicz inequality

In this section, we introduce the Łojasiewicz inequality tool as an abstract convergence theorem. We show that one can deduce the convergence of an iteration algorithm from a triplet of conditions (L), (D), and (S). Further-

[1] This property is nontrivial. Although an eigenstate of the nonlinear problem is always an eigenstate of the linearized problem, it is not always the *lowest energy* eigenstate (i.e., ground state) of the linearized problem.

[2] Both *exponential* and *linear convergence* refer to the case where $\mathrm{err}_k \leq c^k \cdot \mathrm{err}_0$ for some $0 < c < 1$. We sometimes use both terms interchangeably. The term *linear convergence* is more popular in the optimization community.

more, whether the convergence rate is exponential (linear) or polynomial (sublinear) is determined by the exponent in the (L) inequality.

Theorem 7.2.1. *Assume that the domain $\mathcal{M}$ is a Hilbert manifold. Let $\|\cdot\|_X$ be a norm on $\mathcal{TM}$, the tangent bundle of $\mathcal{M}$, and $\|\cdot\|_Y$ be a norm in the ambient space of $\mathcal{M}$ which is complete. Here $\|\cdot\|_X$ and $\|\cdot\|_Y$ can be either same or different. Let $\{u_n\}_{n=0}^{\infty} \subset \mathcal{M}$ be a sequence generated by some iterative algorithm. Assume that $E(u)$ is differentiable on $\mathcal{M}$ and let grad $E(u)$ be the manifold gradient of $E(u)$. If $E(u)$ and $\{u_n\}_{n=0}^{\infty}$ satisfy the following conditions for all $n \in \mathbb{Z}_+$:*

- (Łojasiewicz Gradient Inequality) *There exists u^* that is a cluster point of $\{u_n\}$, and there exists $0 < C_L < +\infty$, $0 < \theta \leq \frac{1}{2}$, such that for large enough n,*
$$|E(u_n) - E(u^*)|^{1-\theta} \leq C_L \|grad\ E(u_n)\|_X; \tag{L}$$
- (Descent Inequality) *There exists $C_D > 0$ such that for large enough n,*
$$E(u_n) - E(u_{n+1}) \geq C_D \|grad\ E(u_n)\|_X \|u_{n+1} - u_n\|_Y; \tag{D}$$
- (Step-size Condition) *There exists $C_S > 0$ such that for large enough n,*
$$\|u_{n+1} - u_n\|_Y \geq C_S \|grad\ E(u_n)\|_X. \tag{S}$$

Then u^* is the unique limit point of $\{u_n\}_{n=0}^{\infty}$ w.r.t. $\|\cdot\|_Y$. Moreover, $\{u_n\}_{n=0}^{\infty}$ converge to u^* with the following asymptotic convergence rate:

$$\|u_n - u^*\|_Y \lesssim \begin{cases} e^{-cn}, & \text{if } \theta = \frac{1}{2}, \\ n^{-\frac{\theta}{1-2\theta}}, & \text{if } \theta \in (0, \frac{1}{2}), \end{cases}$$

where $c := log\,(1 - \frac{C_D C_S}{2C_L^2})$.

Proof. $\{E(u_n)\}$ is monotonically decreasing from Condition (D). Since u^* is a cluster point of $\{u_n\}$, $E(u_n) \geq E(u^*)$ for any n. We also have $\lim_{n\to\infty} E(u_n) = E(u^*)$ by continuity of $E(\cdot)$. Without loss of generality, assume that $E(u^*) = 0$. By Conditions (D) and (L), we have

$$\begin{aligned} \|u_{n+1} - u_n\|_Y &\leq \frac{E(u_n) - E(u_{n+1})}{C_D \|\text{grad } E(u_n)\|_X} \leq \frac{C_L}{C_D}(E(u_n) - E(u_{n+1}))E(u_n)^{\theta-1} \\ &\leq \frac{C_L}{C_D} \int_{E(u_{n+1})}^{E(u_n)} y^{\theta-1}\,\mathrm{d}y = \frac{C_L}{\theta C_D}(E(u_n)^{\theta} - E(u_{n+1})^{\theta}). \end{aligned}$$

Using a bootstrapping argument, we have that for any $m > n$,

$$\|u_n - u_m\|_Y \le \frac{C_L}{\theta C_D}(E(u_n)^\theta - E(u_m)^\theta) \le \frac{C_L}{\theta C_D}E(u_n)^\theta. \tag{7.4}$$

Since $E(u_n)$ is convergent, we deduce that u_n is convergent, and the limit point is u^*.

To estimate the convergence rate, let $r_n := \sum_{k=n}^\infty \|u_{k+1}-u_k\|_Y$, then $\|u_n-u^*\|_Y \le r_n$. It suffices to estimate the convergence rate of r_n. By Conditions (L) and (S), for large enough n,

$$|E(u_n) - E(u^*)|^{1-\theta} \le C_L\|\text{grad } E(u_n)\|_X \le \frac{C_L}{C_S}\|u_{n+1} - u_n\|_Y.$$

Since we have made the assumption that $E(u^*) = 0$, we obtain

$$E(u_n) \le \left(\frac{C_L}{C_S}\|u_{n+1} - u_n\|_Y\right)^{\frac{1}{1-\theta}}. \tag{7.5}$$

Thus, we have

$$\begin{aligned} r_n = \sum_{k=n}^\infty \|u_{k+1} - u_k\|_Y &\le \sum_{k=n}^\infty \frac{C_L}{\theta C_D}(E(u_k)^\theta - E(u_{k+1})^\theta) = \frac{C_L}{\theta C_D}E(u_n)^\theta \\ &\le \frac{C_L}{\theta C_D}\left(\frac{C_L}{C_S}\|u_{n+1} - u_n\|_Y\right)^{\frac{\theta}{1-\theta}} = \frac{C_L}{\theta C_D}\left(\frac{C_L}{C_S}(r_n - r_{n+1})\right)^{\frac{\theta}{1-\theta}}, \end{aligned}$$

where the first inequality is due to (7.4) and the second inequality is due to (7.5). This gives

$$r_{n+1} \le r_n - Cr_n^{\frac{1-\theta}{\theta}}, \quad C := C_L^{-\frac{1}{\theta}}(\theta C_D)^{\frac{1-\theta}{\theta}}C_S.$$

Note that here $0 < C < 1$, otherwise the sequence would have converged in finite steps.

If $\theta \in (0, \frac{1}{2})$, let $s_n := s_0 n^{-\gamma}$, $\gamma = \frac{\theta}{1-2\theta}$, and $s_0 \ge \max\{r_0, (C/\gamma)^{-\gamma}\}$. Then

$$s_{n+1} = s_n\left(1 + \frac{1}{n}\right)^{-\gamma} \ge s_n\left(1 - \frac{1}{n}\cdot\gamma\right) = s_n\left(1 - \gamma s_0^{-1/\gamma}s_n^{1/\gamma}\right) \ge s_n - Cs_n^{\frac{\gamma+1}{\gamma}} = s_n - Cs_n^{\frac{1-\theta}{\theta}}.$$

Combining $s_0 \ge r_0$, $r_{n+1} \le r_n - Cr_n^{\frac{1-\theta}{\theta}}$, and $s_{n+1} \ge s_n - Cs_n^{\frac{1-\theta}{\theta}}$, by induction,

$$r_n \le s_n = s_0 n^{-\frac{\theta}{1-2\theta}} \quad \forall n,$$

which is polynomial (or sub-linear) convergence.

If $\theta = \frac{1}{2}$, then $r_{n+1} \leq (1-C)r_n$, and

$$r_n \leq r_0 e^{cn}, \quad c := \ln(1-C),$$

which is exponential (or linear) convergence. □

The above result can be seen as a generalization of Theorem 2.3 in [118] to the Hilbert space/manifold. Another work in this direction is [59]. What is new in our version is that one has the freedom to choose mixed norms ($\|\cdot\|_X$ and $\|\cdot\|_Y$), as long as the conditions (L), (D), and (S) can be satisfied under these norms. One example is the $\|\cdot\|_{a_u}$ in this chapter, which varies with u.

The advantage of the Łojasiewicz inequality approach is that instead of dealing with the time discretization of the gradient flow, it gives the convergence of the gradient descent directly. The triplet of conditions (L), (D), and (S) in Theorem 7.2.1 all have clear and intuitive meanings. In fact, it is easier to deduce the convergence property of the gradient flow from that of the gradient descent, since we only need to take the limit $\tau \to 0^+$; while the reverse direction from gradient flow to gradient descent can be more difficult.

An important observation is that the exponent θ in Łojasiewicz gradient inequality indicates the *degree of polynomial* of the objective function. For example, consider $x \in \mathbb{R}$, let $f(x) = x^k$ for a positive integer k, then Łojasiewicz gradient inequality holds with $\theta = 1/k$. From this viewpoint, exponential convergence is closely related to certain quadratic-like behavior of the objective functional. It is thus unusual for a quartic-quadratic functional $E(\cdot)$ (i.e., a functional which is the sum of nonnegative quartic and quadratic terms) to have exponential convergence rate. What the Sobolev gradient does is to force the quartic term to behave like quadratic. This is the idea behind the proof of Theorem 7.3.9.

7.3 Exponential convergence of Sobolev gradient descent

In this section, we establish the convergence rate of the a_u-Sobolev gradient descent for Problems (7.1) and (7.2). In Section 7.3, we introduce the setting of manifold optimization and derive the a_u-Sobolev gradient descent method. In Section 7.3, using the Łojasiewicz inequality tool from the previous section, we prove the exponential convergence rate by checking conditions (L), (D) and (S) for this specific method.

Manifold setting and derivation of a_u-Sobolev gradient descent.

The following assumptions on Ω, V and β will be required throughout this section.

Assumptions 7.3.1. *Let Ω, V and β be chosen such that the following assumptions hold:*

- *Ω is a bounded domain in $\mathbb{R}^d$, $d = 1,\, 2$, or 3, and Ω is either convex Lipschitz or has a smooth boundary;*
- *$V \geq 0$ and $V \in L^\infty(\Omega)$, V is a trapping potential, and $\beta \geq 0$.*

Remark 7.3.2. V is chosen as a trapping potential so that the eigenstates of interest are localized. It is then natural to impose zero Dirichlet boundary conditions on $\partial\Omega$. Examples of a trapping potential include the well model in the classical Anderson localization where $\lim_{|x|\to\infty} V(x) = +\infty$, and the fully disordered model with high contrast and small interaction length.

Define the infinite dimensional Hilbert manifold $\mathcal{M}$ as

$\mathcal{M} := \{u \in H_0^1(\Omega) : \|u\|_{L^2(\Omega)} = 1,\ \|u\|_{L^\infty(\Omega)} \leq M_0$ for some global constant $M_0\}$.

Then $\mathcal{M}$ is a submanifold in $H_0^1(\Omega) \cap L^\infty(\Omega)$. Note that although the original problem (7.1) allows $v(x) \in \mathbb{C}$, we restrict our search to $u(x) \in \mathbb{R}$, as we will see that the existence of a real and positive ground state is ensured by Theorem 7.3.4. We also remark that $\|u\|_{L^\infty(\Omega)} \leq M_0$ is not directly guaranteed by the iterative algorithm, but is rather left as an assumption. It is a plausible assumption because we will see that the ground state v is in $L^\infty(\Omega)$ by Hölder continuity in Theorem 7.3.4.

For simplicity we drop Ω in norm and inner product notations when there is no confusion. The *tangent space* of $\mathcal{M}$ at point $u \in \mathcal{M}$ is defined as

$$\mathcal{T}_u\mathcal{M} = \{\xi \in H_0^1(\Omega) \cap L^\infty(\Omega) :\ (\xi, u)_{L^2} = 0\}. \tag{7.6}$$

We need an inner product in the tangent space, denoted as $(\cdot,\cdot)_X$. On the finite dimensional Riemannian manifold, this is dubbed the *Riemannian metric*. It can be easily generalized to the infinite dimensional Hilbert manifold.

For $u \neq 0$, the *retraction* of u onto $\mathcal{M}$ is given by

$$R(u) = u/\|u\|_{L^2}.$$

Note that the retraction operation itself is independent of the choice of the inner product $(\cdot,\cdot)_X$, but its approximation property is not. When the inner product $(\cdot,\cdot)_X$ is introduced, it is usually required that the retraction is at least first-order, i.e., $R(z+\xi) = z + o(\|\xi\|_X)$ for $z \in \mathcal{M}$ and $\xi \in \mathcal{T}_u\mathcal{M}$.

Given an inner product $(\cdot,\cdot)_X$, let $\mathcal{G}$ be its associated Greens operator, i.e.,

$$(z, \mathcal{G}w)_X = (z, w)_{L^2}, \qquad \forall\, z, w \in X.$$

For an arbitrary element ξ in the ambient space, the *projection onto the tangent space* at point $u \in \mathcal{M}$ is given by

$$P_{\mathcal{T}_u\mathcal{M}}(\xi) = \xi - \frac{(\xi, u)_{L^2}}{(\mathcal{G}u, \mathcal{G}u)_X}\mathcal{G}u.$$

Given a differentiable function $E(u)$ defined on $\mathcal{M}$, the *Sobolev gradient* of $E(u)$ with respect to the inner product $(\cdot,\cdot)_X$ is the unique element $\nabla_X E(u) \in X$ such that

$$(\nabla_X E(u), w)_X = (\nabla E(u), w)_{L^2}, \qquad \forall\, w \in X.$$

The *manifold gradient* of $E(u)$ on $\mathcal{M}$, denoted as grad$E(u)$, is the projection of the Sobolev gradient onto the tangent space with respect to the inner product $(\cdot,\cdot)_X$. Thus we have

$$\operatorname{grad} E(u) = P_{\mathcal{T}_u\mathcal{M}}(\nabla_X E(u)) = \nabla_X E(u) - \frac{(\nabla_X E(u), u)_{L^2}}{(\mathcal{G}u, \mathcal{G}u)_X}\mathcal{G}u.$$

It can be inferred from the above expression that grad $E(u) = 0$ implies $\nabla E(u) = \lambda u$ for some scalar λ. If $E(u)$ is as in (7.2), then u is an eigenstate of (7.1). This fact is independent of the choice of inner product $(\cdot,\cdot)_X$.

The choice of the inner product in the tangent space plays an important role in the analysis of manifold optimization algorithms as different inner products give different forms of gradient flow and gradient descent algorithms. Popular choices include L^2, H^1, and the a_0 inner product defined as follows:

$$(z, w)_{a_0} := \int_\Omega \nabla z \nabla w + V z w, \qquad \forall\, z,\ w \in \mathcal{T}_u\mathcal{M}, \quad u \in \mathcal{M}.$$

All the above inner products are fixed everywhere on the manifold. Things become interesting when the inner product becomes adapted to u. Specifically, we are interested in the following inner product

$$(z, w)_{a_u} := \int_\Omega \nabla z \nabla w + V z w + \beta |u|^2 z w, \qquad \forall\, z,\ w \in \mathcal{T}_u\mathcal{M}, \quad u \in \mathcal{M}, \tag{7.7}$$

and we define

$$\mathcal{A}_u := -\Delta + V + \beta|u|^2, \tag{7.8}$$

such that $(\mathcal{A}_u z, w)_{L^2} = (z, w)_{a_u}$ for any z, w. This new inner product $(\cdot,\cdot)_{a_u}$ can be seen as the linearization of the Gross-Pitaevskii energy functional. A desirable property of this inner product is that the Sobolev gradient of $E(u)$ is u itself, i.e.,

$$\nabla_{a_u} E(u) = u. \tag{7.9}$$

This inner product has the associated Greens operator $\mathcal{G}_u$ whose properties have been explored in [72].

Lemma 7.3.3. *Under the adaptive inner product $(\cdot,\cdot)_{a_u}$, the retraction R is second-order.*

Proof. For $u \in \mathcal{M}$ and for any $\xi \in \mathcal{T}_u\mathcal{M}$,

$$\frac{\|R(u+\xi) - (u+\xi)\|_{a_u}}{\|u+\xi\|_{a_u}} = \frac{\|(1 - 1/\|u+\xi\|_{L^2})(u+\xi)\|_{a_u}}{\|u+\xi\|_{a_u}} = \left|1 - \frac{1}{\|u+\xi\|_{L^2}}\right|.$$

Note that ξ is a tangent vector of the manifold at u. By (7.6), $\|u+\xi\|_{L^2}^2 = \|u\|_{L^2}^2 + \|\xi\|_{L^2}^2 + 2(\xi, u)_{L^2} = 1 + \|\xi\|_{L^2}^2$. Thus we have

$$\frac{\|R(u+\xi) - (u+\xi)\|_{a_u}}{\|u+\xi\|_{a_u}} = \left|1 - (1 + \|\xi\|_{L^2}^2)^{-1/2}\right| = \frac{1}{2}\|\xi\|_{L^2}^2 + \mathcal{O}(\|\xi\|_{L^2}^4).$$

By the Poincaré inequality, when $V \geq 0$ and $\beta \geq 0$,

$$\|\xi\|_{L^2}^2 \leq C_P\|\nabla\xi\|_{L^2}^2 \leq C_P\|\xi\|_{a_u}^2$$

for some domain constant $C_P > 0$. Thus we have

$$\|R(u+\xi) - (u+\xi)\|_{a_u} = \mathcal{O}(\|\xi\|_{a_u}^2),$$

where the constant in $\mathcal{O}(\cdot)$ is independent of ξ. □

Using the inner product $(\cdot,\cdot)_{a_u}$, the manifold gradient becomes

$$\text{grad } E(u) = u - \frac{(u,\, u)_{L^2}}{(\mathcal{G}_u u,\, \mathcal{G}_u u)_{a_u}}\mathcal{G}_u u. \tag{7.10}$$

We now have the Sobolev projected gradient descent (Sobolev PGD) as in (7.3):

$$\begin{aligned} u_{n+1} &= R\left(u_n - \tau_n \cdot \text{grad } E(u_n)\right) \\ &= R\left((1-\tau_n)\, u_n + \tau_n \cdot \frac{(u_n, u_n)_{L^2}}{(\mathcal{G}_{u_n} u_n, \mathcal{G}_{u_n} u_n)_{a_{u_n}}}\mathcal{G}_{u_n} u_n\right). \end{aligned} \tag{7.11}$$

Asymptotic convergence and exponential rate.

Throughout the rest of the chapter, let v always denote the global minimizer of $E(u)$, i.e., the ground state of the nonlinear eigenproblem. Let λ always denote its corresponding eigenvalue. We have the following basic observations about the ground state v.

Theorem 7.3.4. *There is a ground state v that satisfies $v(x) > 0$ everywhere on Ω. It is the only strictly positive eigenstate of (7.1) up to scaling. Moreover, it is both the* unique *ground state of the nonlinear eigenproblem (7.1) and the* unique *ground state of the linearized operator $\mathcal{A}_v$ up to the sign. Moreover, v has Hölder regularity $v \in C^{0,\alpha}(\bar{\Omega})$ for some $0 < \alpha < 1$.*

Proof. This theorem is a consequence of Lemma 2 in [28] and Lemmas 5.3 and 5.4 in [72]. We only outline the main idea of the proof here to make this chapter self-contained.

The idea is that the existence of at least one global minimizer v is ensured by the convexity of $E(u)$. The Hölder continuity of v is ensured by elliptic regularity, see e.g., [67, Theorem 8.24]. This v can always be chosen to be nonnegative because $E(u) = E(|u|)$. This nonnegativity can be made into positivity by applying the Harnack inequality to $(A_v - \lambda)$, see e.g., [67, Corollary 8.21]. Thus, there exists a ground state of the nonlinear problem that is positive. The same argument shows that the ground state eigenfunction of the linearized operator $\mathcal{A}_v$ is also positive and is unique. Since v is an eigenfunction of $\mathcal{A}_v$ and is positive, it is exactly that ground state. Thus we have the "double ground state" property. Finally, the uniqueness of any positive eigenstate of the original nonlinear eigenproblem can be established by contradiction. This can be done either by the Picone identity as in [72], or by showing that as long as some u itself is the ground state of the linearized operator $\mathcal{A}_u$, it must be the ground state of the original problem. □

It turns out in subsequent results that v being the "double" ground state in Theorem 7.3.4 is essential to the exponential convergence rate.

Lemma 7.3.5. *If the initial point u_0 of the Sobolev PGD satisfies $u_0 > 0$ everywhere on Ω, then $\{u_n\}_{n=0}^{\infty}$ generated by the Sobolev PGD with step size $\tau_{min} \leq \tau_n \leq \tau_{max}$ for some $0 < \tau_{min} \leq \tau_{max} \leq 1$ converges to the ground state v strongly in $H^1(\Omega)$.*

Proof. The proof is originally developed in [72] and we only outline its main idea here to make this chapter self-contained. The key idea is to show that $u_n(x) \geq 0$ for all n by induction. Assume that $u_n \geq 0$, we will show that this implies $\mathcal{G}_{u_n} u_n \geq 0$, and with $\tau_n \leq 1$ this implies $u_{n+1} \geq 0$.

Specifically, observe that $\mathcal{G}_{u_n} u_n$ is the unique minimizer of

$$\phi(y) := (y, y)_{a_{u_n}} - 2(y, u_n)_{L^2}.$$

Since $u_n \geq 0$, we have that $\phi(|y|) \leq \phi(y)\ \forall y$. This implies that the minimizer of $\phi(\cdot)$ is nonnegative because we can always take the absolute value of the variable without increasing the functional value. Thus, $\mathcal{G}_{u_n} u_n \geq 0$. We then use the fact that u_{n+1} is the scaled weighted average of two nonnegative quantities:

$$\tilde{u}_{n+1} = (1 - \tau_n) u_n + \tau_n \gamma_n \mathcal{G}_{u_n} u_n, \quad \gamma_n = \frac{(u_n, u_n)_{L^2}}{(\mathcal{G}_{u_n} u_n, \mathcal{G}_{u_n} u_n)_{a_{u_n}}} \geq 0, \quad u_{n+1} = \tilde{u}_{n+1} / \|\tilde{u}_{n+1}\|_{L^2}.$$

Thus, we establish that $u_n \geq 0$ implies $u_{n+1} \geq 0$. Since $u_0 > 0$, we have that $u_n \geq 0$ for all n.

The existence of a cluster point u^* for $\{u_n\}$ can be ensured by energy decay. This convergence to u^* is in the sense of weak convergence in $H_0^1(\Omega)$. From the above induction, u^* is nonnegative, and following an argument similar to that in Theorem 7.3.4 we can show that it is all positive.

Since the step size is lower-bounded, u^* must be a fixed point of $E(u)$, where $\operatorname{grad} E(u^*) = 0$. As we mentioned above, $\operatorname{grad} E(u^*) = 0$ implies $\nabla E(u^*) = \lambda u^*$ for some scalar λ, i.e., u^* is an eigenstate of the eigenvalue problem (7.1). From the uniqueness result of positive eigenstate in Theorem 7.3.4 we know that it could only be the ground state v. Therefore, $\{u_n\}$ converges to v itself.

Finally, the weak convergence in $H_0^1(\Omega)$ implies strong convergence in $L^p(\Omega)$ for $p < 6$ by the Rellich-Kondrachov embedding. This would give the convergence of energy $\{E(u_n)\}$, and consequently strong convergence in $H^1(\Omega)$. □

Before proceeding to the proof of Conditions (L), (D), and (S), we first need some technical lemmas.

Lemma 7.3.6 (Norm equivalence)**.** *Under Assumptions 7.3.1, there exist positive constants C_E, $\widetilde{C_E}$ depending only on β, M_0, V, and the domain Ω, such that*

$$
\begin{aligned}
C_E \|\cdot\|_{a_u} &\le \|\cdot\|_{a_0} \le C_E^{-1} \|\cdot\|_{a_u}, \\
\widetilde{C_E} \|\cdot\|_{a_u} &\le \|\cdot\|_{H^1} \le \widetilde{C_E}^{-1} \|\cdot\|_{a_u}.
\end{aligned}
$$

Proof. As for the equivalence between $\|\cdot\|_{a_0}$ and $\|\cdot\|_{a_u}$, the second part of the inequality holds for all $0 < C_E \le 1$ since u^2 is nonnegative. For the first part, by Poincaré inequality, $\|z\|_{L^2}^2 \le C_P |z|_{H^1}^2$ for some domain constant $C_P = C_P(\Omega)$. Thus, we have

$$
\begin{aligned}
\|z\|_{a_0}^2 - C_E \|z\|_{a_u}^2 &= (1 - C_E)|z|_{H^1}^2 + \int_\Omega ((1 - C_E)V - C_E \beta u^2) z^2 \\
&\ge (1 - C_E)|z|_{H^1}^2 - C_E \beta \int_\Omega u^2 z^2 \\
&\ge (1 - C_E - C_E \beta M_0^2 C_P)|z|_{H^1}^2, \qquad \forall z \in H_0^1(\Omega), \quad C_E \le 1.
\end{aligned}
$$

Take $0 < C_E \le 1/(1 + \beta M_0^2 C_P)$, then $C_E \|z\|_{a_u}^2 \le \|z\|_{a_0}^2$.

As for the equivalence between $\|\cdot\|_{a_u}$ and $\|\cdot\|_{H^1}$, we have

$$
\begin{aligned}
\|z\|_{H^1}^2 - \widetilde{C_E} \|z\|_{a_u}^2 &= \|z\|_{H^1}^2 - \widetilde{C_E} |z|_{H^1}^2 - \widetilde{C_E} \int_\Omega (V + \beta u^2) z^2 \\
&\ge \left(1 - \widetilde{C_E} - \widetilde{C_E} C_P (\|V\|_{L^\infty} + \beta M_0^2)\right) |z|_{H^1}^2, \qquad \forall z \in H_0^1(\Omega), \quad \widetilde{C_E} \le 1.
\end{aligned}
$$

Take $0 < \widetilde{C_E} \le 1/(1 + C_P(\|V\|_{L^\infty} + \beta M_0^2))$, then $\widetilde{C_E} \|z\|_{a_u}^2 \le \|z\|_{H^1}^2$. On the other hand,

$$
\begin{aligned}
\widetilde{C_E}^{-1} \|z\|_{a_u}^2 - \|z\|_{H^1}^2 &= (\widetilde{C_E}^{-1} - 1)|z|_{H^1}^2 + \widetilde{C_E}^{-1} \int_\Omega (V + \beta u^2) z^2 - \|z\|_{L^2} \\
&\ge \left(C_P^{-1}(\widetilde{C_E}^{-1} - 1) + \widetilde{C_E}^{-1} \beta M_0^2 - 1\right) \|z\|_{L^2}.
\end{aligned}
$$

Take $0 < \widetilde{C_E} \le (1 + C_P \beta M_0^2)/(1 + C_P)$, then $\|z\|_{H^1}^2 \le \widetilde{C_E}^{-1} \|z\|_{a_u}^2$. The final choice of $\widetilde{C_E}$ is the smaller of the two. □

In the next two lemmas, let λ_i and μ_i be the ith smallest eigenvalues of $\mathcal{A}_v$ and $\mathcal{A}_u$ respectively, and v_i and w_i be their corresponding eigenfunctions satisfying $\|v_i\|_{L^2} = 1$ and $\|w_i\|_{L^2}$ (so that $v = v_1$, $\lambda = \lambda_1$). Theorem 7.3.4 has ensured the uniqueness of the ground state. The fact that $\mathcal{A}_v$ only has point spectrum ensures that there is a positive gap C_v between λ_1 and λ_2.

Lemma 7.3.7 (Perturbation of eigenvalues and eigenfunctions)**.** *Under Assumptions 7.3.1, there exists a positive constant $C = C(\beta, V, M_0, \Omega, \lambda_1, C_v)$, such that for all $u \in \mathcal{M}$ satisfying $\|u - v\|_{H^1} \leq C$, we have that $\|u - w_1\|_{L^2} \leq s$ for some $s < 1$.*

Proof. For notational simplicity, we allow the constants C, C' to change their meanings through the proof. We also denote

$$t := \|u - v\|_{H^1}.$$

Using the variational form of the eigenvalues, we have

$$\begin{aligned}
\mu_1 &= \min_{\substack{z \in H_0^1(\Omega), \\ \|z\|_{L^2}=1}} (z, z)_{a_u} \leq (v, v)_{a_u}, \\
\lambda_1 &= \min_{\substack{z \in H_0^1(\Omega), \\ \|z\|_{L^2}=1}} (z, z)_{a_v} \leq (w_1, w_1)_{a_v}, \\
\lambda_1 + \lambda_2 &= \min_{\substack{z_1, z_2 \in H_0^1(\Omega), \\ \|z_1\|_{L^2}=\|z_2\|_{L^2}=1, \\ z_1 \perp z_2}} (z_1, z_1)_{a_v} + (z_2, z_2)_{a_v} \leq (w_1, w_1)_{a_v} + (w_2, w_2)_{a_v}.
\end{aligned}$$

We will use the above relations to bound the gap between μ_1 and λ_1, and λ_2 and μ_2. First, we have

$$\begin{aligned}
\mu_1 \leq (v, v)_{a_u} &= (v, v)_{a_v} + \int_\Omega \beta(u^2 v^2 - v^4) \\
&= \lambda_1 + \int_\Omega \beta v^2 (u + v)(u - v) \\
&\leq \lambda_1 + 2\beta M_0^3 \int_\Omega |u - v| \\
&\leq \lambda_1 + C(\beta, M_0, \Omega) \cdot t.
\end{aligned}$$

Therefore, there exists $C = C(\beta, M_0, \Omega)$ such that when $t \leq C$,

$$\mu_1 \leq \lambda_1 + \frac{1}{6} C_v. \tag{7.12}$$

Next, we note that

$$\begin{aligned}
\lambda_1 + \lambda_2 &\leq (w_1, w_1)_{a_v} + (w_2, w_2)_{a_v} \\
&= (w_1, w_1)_{a_u} + (w_2, w_2)_{a_u} + \int_\Omega \beta(v^2 - u^2)(w_1^2 + w_2^2) \\
&= \mu_1 + \mu_2 + \int_\Omega \beta(v + u)(v - u)(w_1^2 + w_2^2).
\end{aligned} \tag{7.13}$$

To estimate $\|w_1\|_{L^\infty}$, note that it is the weak solution of

$$-\Delta w_1 + V w_1 + \beta u^2 w_1 = \mu_1 w_1.$$

Since $V,\ u \in L^\infty(\Omega)$, by elliptic regularity, we get

$$\begin{aligned}\|w_1\|_{H^2} &\le C(\beta, V, M_0, \Omega)(\|w_1\|_{H^1} + \mu_1 \|w_1\|_{L^2})\\ &\le C(\beta, V, M_0, \Omega) + C'(\beta, V, M_0, \Omega) \cdot \mu_1.\end{aligned}$$

When $d \le 3$, using Sobolev embedding, we obtain

$$\|w_1\|_{L^\infty} \le C(\Omega)\|w_1\|_{H^2}.$$

Since we have shown that $\mu_1 \le \lambda_1 + C \cdot t$, putting them together we have

$$\|w_1\|_{L^\infty} \le C(\beta, V, M_0, \Omega, \lambda_1) + C'(\beta, V, M_0, \Omega, \lambda_1) \cdot t.$$

Similarly, we can prove that[3]

$$\|w_2\|_{L^\infty} \le C(\beta, V, M_0, \Omega, \lambda_1, \lambda_2) + C'(\beta, V, M_0, \Omega, \lambda_1, \lambda_2) \cdot t.$$

Plugging them back into (7.13), we have

$$(w_1, w_1)_{a_v} + (w_2, w_2)_{a_v} \le \mu_1 + \mu_2 + (C(\beta, V, M_0, \Omega, \lambda_1, \lambda_2) + C'(\beta, V, M_0, \Omega, \lambda_1, \lambda_2) \cdot t)^2 \cdot t.$$

Therefore, there exists $C = C(\beta, V, M_0, \Omega, \lambda_1, \lambda_2)$, such that when $t \le C$,

$$\lambda_1 + \lambda_2 \le \mu_1 + \mu_2 + \frac{1}{6} C_v. \tag{7.14}$$

Combining (7.12) and (7.14), we have

$$\mu_1 \le \lambda_1 + \frac{1}{6} C_v, \qquad \mu_2 \ge \lambda_2 - \frac{1}{3} C_v, \qquad \mu_2 - \mu_1 \ge \frac{1}{2} C_v. \tag{7.15}$$

Next, note that

$$\begin{aligned}\lambda_1 \le (w_1, w_1)_{a_v} &= (w_1, w_1)_{a_u} + \int_\Omega \beta(v^2 - u^2) w_1^2\\ &\le \mu_1 + C(\beta, V, M_0, \Omega)\|w_0\|_{L^\infty}^2 \cdot t\\ &\le \mu_1 + (C(\beta, V, M_0, \Omega) + C'(\beta, V, M_0, \Omega) \cdot t)^2 \cdot t.\end{aligned}$$

[3] We omit the details of showing $\mu_2 \le \lambda_2 + C \cdot t$ by showing $\mu_1 + \mu_2 \le \lambda_1 + \lambda_2 + C \cdot t$ using the variational form.

Therefore, there exists $C = C(\beta, V, M_0, \Omega, \lambda_1)$ such that when $t \leq C$,

$$\lambda_1 \leq \mu_1 + \frac{1}{6} C_v. \tag{7.16}$$

Equations (7.12), (7.15), and (7.16) contain all the relations between λ_1, λ_2, μ_1, and μ_2 that we will need.

Since $\{w_i\}_{i=1}^{\infty}$ forms an orthonormal basis of $H_0^1(\Omega)$, in order to estimate $\|u - w_1\|_{L^2}$, it suffices to bound $(u,u)_{a_u} - \mu_1$. Note that

$$\begin{aligned}
(u,u)_{a_u} - \lambda_1 &= (u,u)_{a_u} - (v,v)_{a_v} \\
&= (u,u)_{a_u} - (v,v)_{a_u} + \int_\Omega \beta(u^2v^2 - v^4) \\
&\leq (\|u\|_{a_u} + \|v\|_{a_u}) \cdot \|u - v\|_{a_u} + \int_\Omega \beta v^2(u+v)(u-v) \\
&\leq C(\beta, V, M_0, \Omega)(\|u\|_{H^1} + \|v\|_{H^1}) \cdot \|u - v\|_{H^1} + \int_\Omega \beta v^2(u+v)(u-v) \\
&\leq C(\beta, V, M_0, \Omega) \cdot t.
\end{aligned}$$

The fourth inequality uses the norm equivalence in Lemma 7.3.6. Thus, there exists $C = C(\beta, V, M_0, \Omega)$, such that when $t \leq C$,

$$(u,u)_{a_u} - \lambda_1 \leq \frac{1}{12} C_v. \tag{7.17}$$

Combining (7.15), (7.16), and (7.17), we have

$$(u,u)_{a_u} - \mu_1 \leq \frac{1}{4} C_v \leq \frac{1}{2}(\mu_2 - \mu_1).$$

Assume that $u = \sum_{i=1}^{\infty} c_i w_i$, where $\sum_{i=1}^{\infty} c_i^2 = 1$. Then we get

$$(u,u)_{a_u} - \mu_1 = \sum_{i=1}^{\infty} c_i^2 \mu_i - \mu_1 \geq c_1^2 \mu_1 + \sum_{i=2}^{\infty} c_i^2 \mu_2 - \mu_1 = (1 - c_1^2)(\mu_2 - \mu_1).$$

Since $(u,u)_{a_u} - \mu_1 \leq \frac{1}{2}(\mu_2 - \mu_1)$, we have

$$1 - c_1^2 \leq \frac{1}{2}, \qquad |c_1| \geq \frac{1}{\sqrt{2}}.$$

If $c_1 \leq -1/\sqrt{2}$, we can use $-w_1$ to replace w_1. Thus, we always have $c_1 \geq 1/\sqrt{2}$. This gives

$$\|u - w_1\|_{L^2} = \sqrt{2 - 2c_1} \leq \sqrt{2 - \sqrt{2}} < 1.$$

In other words, $s \leq \sqrt{2 - \sqrt{2}}$. The constant C in the statement of the lemma is the smallest of all the constants C, C' in the proof. Since $\lambda_2 = \lambda_1 + C_v$, the dependence on λ_2 is the dependence on C_v. □

Lemma 7.3.8 (Condition (L) for the linearized operator)**.** *Let $\mathcal{A} : X \to X$ be a symmetric and positive definite linear operator on the Hilbert space with a bounded Greens operator $\mathcal{G}$. Let μ_i denote the i-th smallest eigenvalue of $\mathcal{A}$, and w_i be its corresponding (normalized) eigenfunction. Assume that $\mu_2 > \mu_1$. Then for any u such that $\|u\|_{L^2} = 1$ and $\|u - w_1\|_{L^2} \le s < 1$, we have*

$$(u, u)_{\mathcal{A}} - (w_1, w_1)_{\mathcal{A}} \le C_L \left((u, u)_{\mathcal{A}} - \frac{1}{(u, \mathcal{G}u)_{L^2}} \right)$$

for some constant C_L that depends only on s, μ_1, and μ_2.

Proof. Since μ_2 is strictly greater than μ_1, we can split $\mathcal{A}$ and u as

$$\begin{aligned} &\mathcal{A} = \mathcal{A}^{(1)} + \mathcal{A}^{(2)}, \quad \mathcal{A}^{(1)} = \mathcal{A}P_{w_1}, \quad \mathcal{A}^{(2)} = \mathcal{A}P_{w_1}^{\perp}, \\ &u = u^{(1)} + u^{(2)}, \quad u^{(1)} = P_{w_1}u, \quad u^{(2)} = P_{w_1}^{\perp}u. \end{aligned}$$

Here P_{w_1} is the orthogonal projection onto the subspace of w_1 under the L^2 inner product, and $P_{w_1}^{\perp} = id - P_{w_1}$. Then $\mathcal{A}^{(1)}u^{(1)} = \mu_1 u^{(1)}$, and $(u^{(2)}, u^{(2)})_{\mathcal{A}^{(2)}} \ge \mu_2 \|u^{(2)}\|_{L^2}^2$ since $u^{(2)} \perp w_1$. By definition of $\mathcal{G}$, $(u, \mathcal{G}v)_{\mathcal{A}} = (u, v)_{L^2}$ for any $u, v \in X$. We have

$$\begin{aligned} (u, \mathcal{G}u^{(1)})_{L^2} &= \mu_1^{-1} \|u^{(1)}\|_{L^2}^2, \\ (u, \mathcal{G}u^{(2)})_{L^2} &= (u^{(1)}, \mathcal{G}u^{(2)})_{L^2} + (u^{(2)}, \mathcal{G}u^{(2)})_{L^2} = (u^{(2)}, \mathcal{G}u^{(2)})_{L^2}, \\ (u^{(2)}, \mathcal{G}u^{(2)})_{L^2} &= (\mathcal{G}u^{(2)}, \mathcal{G}u^{(2)})_{\mathcal{A}} \ge \mu_2 \|\mathcal{G}u^{(2)}\|_{L^2}^2 \\ &= \mu_2 \|u^{(2)}\|_{L^2}^{-2} \cdot (\|\mathcal{G}u^{(2)}\|_{L^2}^2 \|u^{(2)}\|_{L^2}^2) \ge \mu_2 \|u^{(2)}\|_{L^2}^{-2} \cdot (u^{(2)}, \mathcal{G}u^{(2)})_{L^2}^2, \\ &\text{i.e., } (u, \mathcal{G}u^{(2)})_{L^2} \le \mu_2^{-1} \|u^{(2)}\|_{L^2}^2. \end{aligned}$$

Therefore, the objective inequality is transformed into

$$\begin{aligned} &C_L \left((u, u)_{\mathcal{A}} - \frac{1}{(u, \mathcal{G}u)_{L^2}} \right) - ((u, u)_{\mathcal{A}} - (w_1, w_1)_{\mathcal{A}}) \\ &= (C_L - 1)(u, u)_{\mathcal{A}} - \frac{C_L}{(u, \mathcal{G}u)_{L^2}} + \mu_1 \\ &= (C_L - 1)((u^{(1)}, u^{(1)})_{\mathcal{A}^{(1)}} + (u^{(2)}, u^{(2)})_{\mathcal{A}^{(2)}}) - \frac{C_L}{(u, \mathcal{G}u^{(1)})_{L^2} + (u, \mathcal{G}u^{(2)})_{L^2}} + \mu_1 \\ &\ge (C_L - 1)(\mu_1 \|u^{(1)}\|_{L^2}^2 + \mu_2 \|u^{(2)}\|_{L^2}^2) - \frac{C_L}{\mu_1^{-1} \|u^{(1)}\|_{L^2}^2 + \mu_2^{-1} \|u^{(2)}\|_{L^2}^2} + \mu_1 \\ &= (C_L - 1)(\mu_1 + (\mu_2 - \mu_1)\|u^{(2)}\|_{L^2}^2) - \frac{C_L \mu_1 \mu_2}{\mu_2 + (\mu_1 - \mu_2)\|u^{(2)}\|_{L^2}^2} + \mu_1 \\ &= (\mu_2 - \mu_1) \frac{((C_L - 1)\mu_2 - C_L \mu_1)\|u^{(2)}\|_{L^2}^2 - (C_L - 1)(\mu_2 - \mu_1)\|u^{(2)}\|_{L^2}^4}{\mu_2 + (\mu_1 - \mu_2)\|u^{(2)}\|_{L^2}^2}. \end{aligned}$$

We look for C_L and u such that the above is greater than or equal to 0. In fact, for any $C_L > 1$, if

$$0 \leq \|u^{(2)}\|_{L^2}^2 \leq \frac{(C_L - 1)\mu_2 - C_L\mu_1}{(C_L - 1)(\mu_2 - \mu_1)},$$

then this is satisfied. Note that $\|u - v_1\|_{L^2} \leq s$ implies $\|u^{(2)}\|_{L^2}^2 \leq s^2$. So the requirement on C_L is

$$C_L \geq 1 + \frac{\mu_2}{(\mu_2 - \mu_1)(1 - s^2)}.$$

□

Using the above technical lemmas, we are now ready to prove the following theorems. They show that the sequence $\{u_n\}$ generated by (7.3) satisfies Conditions (L), (D), and (S).

The first theorem is on Condition (L) near the ground state v of the nonlinear eigenproblem. It is the central one of the three theorems.

Theorem 7.3.9. *Under Assumptions 7.3.1, Condition (L) is satisfied for $\|\cdot\|_X = \|\cdot\|_{a_u}$ and $\theta = \frac{1}{2}$ near the ground state v. In other words, there exists some constant $C > 0$, such that for any u in $\{u : u \in \mathcal{M},\ E(u) \geq E(v),\ \|u - v\|_{H^1} \leq C\}$, we have*

$$|E(u) - E(v)|^{\frac{1}{2}} \leq C_L \|grad\ E(u)\|_{a_u}.$$

Proof. First notice that for any u in the constraint set of the theorem, $E(u) - E(v) \leq a_u(u, u) - a_u(v, v)$. This is because

$$\begin{aligned} E(u) - E(v) - \big((u,u)_{a_u} - (v,v)_{a_u}\big) &= -\frac{\beta}{2}\int_\Omega u^4 - \frac{\beta}{2}\int_\Omega v^4 + \beta\int_\Omega u^2v^2 \\ &= -\frac{\beta}{2}\int_\Omega (u^2 - v^2)^2 \leq 0. \end{aligned}$$

Let w_1 be the eigenfunction corresponding to the smallest eigenvalue of $\mathcal{A}_u$, then

$$(u,u)_{a_u} - (v,v)_{a_u} \leq (u,u)_{a_u} - (w_1,w_1)_{a_u}.$$

On the other hand, by (7.10), we have

$$\|\text{grad } E(u)\|_{a_u}^2 = \left\| u - \frac{(u,\, u)_{L^2}}{(\mathcal{G}_u u,\, \mathcal{G}_u u)_{a_u}}\mathcal{G}_u u \right\|_{a_u}^2 = \left\| u - \frac{\mathcal{G}_u u}{(u, \mathcal{G}_u u)_{L^2}} \right\|_{a_u}^2 = (u,u)_{a_u} - \frac{1}{(u, \mathcal{G}_u u)_{L^2}}.$$

It suffices to show that

$$(u,u)_{a_u} - (w_1,w_1)_{a_u} \le C_L \left((u,u)_{a_u} - \frac{1}{(u,\mathcal{G}_u u)_{L^2}} \right), \tag{7.18}$$

which only involves the inner product $(\cdot,\cdot)_{a_u}$.

Using Lemma 7.3.7, we have that there exists $C > 0$ such that when $\|u - v\|_{H^1} < C$, we have $\|u - w_1\|_{L^2} \le s$ for some constant $s < 1$. Thus, Lemma 7.3.8 is applicable to $(\cdot,\cdot)_{a_u}$. This gives the above inequality on $(\cdot,\cdot)_{a_u}$, with a constant C_L depending only on β, V, M_0, Ω, λ_1, and C_v. The Łojasiewicz inequality can thus be achieved. □

Remark 7.3.10. The above proof of Condition (L) depends crucially on Lemma 7.3.8. Lemma 7.3.8 can be seen as the version of the Łojasiewicz inequality with $\theta = \frac{1}{2}$ for a linear operator $\mathcal{A}$. So its primary consequence is the linear convergence rate of the proposed algorithm to the ground state of a linear operator $\mathcal{A}$.

The key idea of the proof Theorem 7.3.9, then, is to reduce it to the inequality (7.18). The inequality (7.18) only involves the operator $\mathcal{A}_u$, which is bilinear. Although $\mathcal{A}_u$ formally depends on u, the inequality (7.18) itself is not affected by nonlinearity. So Lemma 7.3.8 can be applied to prove (7.18).

Thus, one way to interpret the proof of Theorem 7.3.9 is to view it as linearizing the nonlinear eigenproblem (7.1) using the adaptive inner product $(\cdot,\cdot)_{a_u}$, so that it preserves the Łojasiewicz property with $\theta = \frac{1}{2}$.

The next theorem is on Condition (D) for the sequence generated by the proposed algorithm.

Theorem 7.3.11. *Under Assumptions 7.3.1, Condition (D) is satisfied for $\|\cdot\|_X = \|\cdot\|_{a_u}$, $\|\cdot\|_Y = \|\cdot\|_{a_0}$ if $\{u_n\}$ is generated by the Sobolev projected gradient descent with step size $0 < \tau_n \le \tau_{max}$ for some $\tau_{max} > 0$, i.e.,*

$$E(u_n) - E(u_{n+1}) \ge C_D \|grad\ E(u_n)\|_{a_{u_n}} \|u_n - u_{n+1}\|_{a_0}.$$

Proof. It is obvious that $\|u_n - u_{n+1}\|_{a_0} \le \|u_n - u_{n+1}\|_{a_{u_n}}$. Since $\{u_n\}$ is generated by the Sobolev projected gradient descent algorithm, we have

$$u_{n+1} = R\left(u_n - \tau_n \cdot \text{grad}\ E(u_n)\right),$$

$$\text{grad}\ E(u_n) = u_n - \frac{(u_n,u_n)_{L^2}}{(\mathcal{G}_{u_n}u_n, \mathcal{G}_{u_n}u_n)_{a_{u_n}}}\mathcal{G}_{u_n}u_n = u_n - \frac{\mathcal{G}_{u_n}u_n}{(u_n,\mathcal{G}_{u_n}u_n)_{L^2}}.$$

The second-order retraction property implies that

$$u_n - u_{n+1} = \tau_n \left(u_n - \frac{\mathcal{G}_{u_n} u_n}{(u_n, \mathcal{G}_{u_n} u_n)_{L^2}} \right) + \mathcal{O}(\tau_n^2).$$

Thus, we obtain

$$\begin{aligned} E(u_n) - E(u_{n+1}) &= \left(u_n - u_{n+1},\ \nabla_{a_{u_n}} E(u_n)\right)_{a_{u_n}} + \mathcal{O}(\|u_n - u_{n+1}\|^2) \\ &= (u_n - u_{n+1},\ u_n)_{a_{u_n}} + \mathcal{O}(\|u_n - u_{n+1}\|^2) \\ &= \tau_n \left(u_n - \frac{\mathcal{G}_{u_n} u_n}{(u_n, \mathcal{G}_{u_n} u_n)_{L^2}},\ u_n \right)_{a_{u_n}} + \mathcal{O}(\tau_n^2) \\ &= \tau_n \left((u_n, u_n)_{a_{u_n}} - \frac{1}{(u_n, \mathcal{G}_{u_n} u_n)_{L^2}} \right) + \mathcal{O}(\tau_n^2). \end{aligned}$$

On the other hand, we have

$$\|\text{grad } E(u_n)\|_{a_{u_n}} = \left((u_n, u_n)_{a_{u_n}} - \frac{1}{(u_n, \mathcal{G}_{u_n} u_n)_{L^2}} \right)^{\frac{1}{2}},$$

and

$$\begin{aligned} \|u_n - u_{n+1}\|_{a_{u_n}} &= \tau_n \left\| u_n - \frac{\mathcal{G}_{u_n} u_n}{(u_n, \mathcal{G}_{u_n} u_n)_{L^2}} \right\|_{a_{u_n}} + \mathcal{O}(\tau_n^2) \\ &= \tau_n \left((u_n, u_n)_{a_{u_n}} - \frac{1}{(u_n, \mathcal{G}_{u_n} u_n)_{L^2}} \right)^{\frac{1}{2}} + \mathcal{O}(\tau_n^2). \end{aligned}$$

This implies that

$$\|\text{grad } E(u_n)\|_{a_{u_n}} \|u_n - u_{n+1}\|_{a_0} \leq \tau_n \left((u_n, u_n)_{a_{u_n}} - \frac{1}{(u_n, \mathcal{G}_{u_n} u_n)_{L^2}} \right) + \mathcal{O}(\tau_n^2).$$

Therefore, there exists a $\tau_{\max} > 0$ such that when $\tau \leq \tau_{\max}$, there exists C_D such that Condition (D) holds. This C_D only depends on $\tau_{\max}$, but is independent of u_n. □

Next, we have the theorem is on Condition (S) for the sequence generated by the proposed algorithm.

Theorem 7.3.12. *Under Assumptions 7.3.1, Condition (S) is satisfied for for* $\|\cdot\|_X = \|\cdot\|_{a_u}$, $\|\cdot\|_Y = \|\cdot\|_{a_0}$ *if* $\{u_n\}$ *is generated by the Sobolev projected gradient descent with step size* $0 < \tau_{min} \leq \tau_n \leq \tau_{max}$ *for some* $0 < \tau_{min} \leq \tau_{max}$*, i.e.,*

$$\|u_{n+1} - u_n\|_{a_0} \geq C_S \|grad\ E(u_n)\|_{a_{u_n}}.$$

Proof. By Lemma 7.3.6, we have $\|u_{n+1} - u_n\|_{a_0} \geq C_E \|u_{n+1} - u_n\|_{a_{u_n}}$ for some constant C_E. Note that in the previous proof we have shown that

$$\|\text{grad } E(u_n)\|_{a_{u_n}} = \left((u_n, u_n)_{a_{u_n}} - \frac{1}{(u_n, \mathcal{G}_{u_n} u_n)_{L^2}} \right)^{\frac{1}{2}}$$

and

$$\|u_n - u_{n+1}\|_{a_{u_n}} = \tau_n \left((u_n, u_n)_{a_{u_n}} - \frac{1}{(u_n, \mathcal{G}_{u_n} u_n)_{L^2}} \right)^{\frac{1}{2}} + O(\tau_n^2).$$

Therefore, when $\tau_{\min} \leq \tau_n \leq \tau_{\max}$ for some $0 < \tau_{\min} \leq \tau_{\max}$, there exists a constant C_S depending only on C_E, $\tau_{\min}$ and $\tau_{\max}$, such that

$$\|u_{n+1} - u_n\|_{a_0} \geq C_S \|\text{grad } E(u_n)\|_{a_{u_n}}.$$

□

Finally, we deduce the following results on the exponential convergence.

Theorem 7.3.13 (Convergence rate of Sobolev PGD). *If the Sobolev projected gradient descent for $E(u)$ converges to the ground state v, and the step size $\{\tau_n\}$ satisfies $0 < \tau_{min} \leq \tau_n \leq \tau_{max}$, then it converges in the a_0-norm with an asymptotic exponential convergence rate.*

Proof. The proof follows directly from Theorems 7.2.1, 7.3.9, 7.3.11, and 7.3.12. □

Theorem 7.3.14 (Global convergence to ground state). *If the initial state u_0 satisfies $u_0 \geq 0$ everywhere on Ω, and the step size $\{\tau_n\}$ satisfies $0 < \tau_{min} \leq \tau_n \leq \tau_{max}$, then the Sobolev projected gradient descent for $E(u)$ converges in the a_0-norm to the unique ground state with an asymptotic exponential convergence rate.*

Proof. Since the initial state is nonnegative, Lemma 7.3.5 ensures the strong convergence of $\{u_n\}$ to the ground state v in $H_0^1(\Omega)$. By Theorem 7.3.13, the asymptotic convergence rate in the a_0-norm is exponentially fast. □

Note that since the domain Ω is bounded, this convergence rate in the a_0-norm implies the exponential convergence rate in the H^1 or L^2 norm. We also remark that the optimal step size with theoretical guarantee depends on the values $\tau_{\min}$ and $\tau_{\max}$, which in turn depend on some properties of the ground state that are not known beforehand, but some practical choices of τ are demonstrated in the numerical experiments in Section 7.6.

7.4 Spatial discretization

To solve the eigenproblem numerically using the computational procedure in the previous sections, we need to discretize the problem in the spatial domain Ω. Let Ω_h be a spatial discretization with grid size h. Note that we only require Ω_h to be a convergent discretization, i.e., the solution to the discrete problem converges to that of the continuous problem as $h \to 0^+$, and the following analysis applies to general discretization schemes. The discretized problem can be written as

$$\min_{\|u_h\|_{L_h^2}=1,\ u_h\in\mathbb{R}^N} E_h(u_h) := \|u_h\|_{\mathcal{L}_h}^2 + \|u_h\|_{V_h}^2 + \frac{\beta}{2}\|u_h\|_{L_h^4}^4, \tag{7.19}$$

where

$$\|u_h\|_{\mathcal{L}_h}^2 = u_h^\top(-\mathcal{L}_h)u_h \cdot h^d, \quad \|u_h\|_{V_h}^2 = \sum_{i=1}^N V_h(i)u_h(i)^2h^d, \quad \|u_h\|_{L_h^p}^p = \sum_{i=1}^N u_h(i)^p h^d.$$

Here N denotes the total number of grid points, (i) is an indexing of the grid points, i.e., $u_h(i)$ is the i-th entry of the vectorized u_h, d is the dimension of the physical space, and $\mathcal{L}_h$ is the discretized Laplacian. The linearized operator $\mathcal{A}_{u,h}$ now has a matrix representation in $\mathbb{R}^{N\times N}$:

$$\mathcal{A}_{u,h} = -\mathcal{L}_h + \operatorname{diag}\{V_h + \beta u_h^{[2]}\},$$

where $u_h^{[2]}(i) := u_h(i)^2$, i.e., $u_h^{[2]}$ is the componentwise squared vector of u_h. The respective norm is defined as $\|y\|_{\mathcal{A}_{u,h}}^2 := y^\top\mathcal{A}_{u,h}y$. We have the following results.

Theorem 7.4.1 (Discrete version of Theorem 7.3.4)**.** *There is a ground state v_h of the discretized problem that satisfies $v_h > 0$ everywhere on Ω_h. It is the unique positive eigenstate of (7.19). Moreover, it is both the* unique *ground state of the nonlinear eigenproblem (7.19) and the* unique *ground state of the linearized operator $\mathcal{A}_{v,h}$ up to the sign.*

Proof. The existence of the ground state follows from the compactness of the constraint set $\{u_h : u_h \in \mathbb{R}^N, \|u_h\|_{L_h^2} = 1\}$ and the boundedness of $E_h(u_h)$. Thus it suffices to prove its uniqueness and positivity. The proofs for the continuous version, i.e., Lemma 2 in [27] and Lemmas 5.3 and 5.4 in [72], need to be slightly modified to suit the discrete case. This is because the

Harnack inequality and the Picone identity are only valid for continuous functions, and we need to establish their discrete counterparts.

One way to do this is to look at the convergence of the discretized eigenvector to its continuous counterpart at the small grid size limit $h \to 0^+$, see e.g., [88]. This is always possible no matter what kind of discretization we use. We do not present the details here.

Another way is to observe that the discretized Laplacian, $\mathcal{L}_h$, is an M-matrix[4] under some typical discretizations. Examples include finite difference discretization, and some P1-finite element discretizations. When $\mathcal{L}_h$ is an M-matrix, the proof can be simplified and the small h constraint can be released. In this case, the proof takes the following steps:

(1) *For any $\mathcal{A}_{u,h}$, its eigenvector corresponding to the smallest eigenvalue can be chosen to be all positive, and is unique up to the sign.*

Since $-\mathcal{L}_h$ has positive diagonals and non-positive off-diagonals, so does $\mathcal{A}_{u,h}$. Let y be the ground state eigenvector of $\mathcal{A}_{u,h}$, then $|y|^\top \mathcal{A}_{u,h}|y| \leq y^\top \mathcal{A}_{u,h} y$. This is because $\mathcal{A}_{u,h}(i,i)y(i)^2 = \mathcal{A}_{u,h}(i,i)|y(i)|^2$ for any $1 \leq i \leq N$, and $\mathcal{A}_{u,h}(i,j)y(i)y(j) \geq \mathcal{A}_{u,h}(i,j)\cdot |y(i)||y(j)|$ for any $i \neq j$. As y is the ground state eigenvector, this implies $y = |y|$, i.e., y is nonnegative. We now show that y is all positive. If this is not true, then y has some positive and some zero entries. So we can always find a zero entry $y(i)$ that is spatially next to a nonzero one, say $y(j)$, i.e., $y(i) = 0$, $y(j) > 0$, and $-\mathcal{L}_h(i,j) < 0$. Then

$$\begin{aligned} 0 &= \lambda y(i) = (\mathcal{A}_{u,h}y)(i) = (-\mathcal{L}_h y)(i) + V_h(i)y(i) + \beta y(i)^3 \\ &= (-\mathcal{L}_h y)(i) = \sum_k -\mathcal{L}_h(i,k)y(k) = \sum_{k \neq i} -\mathcal{L}_h(i,k)y(k) \leq -\mathcal{L}_h(i,j)y(j) < 0, \end{aligned}$$

which is a contradiction. Thus y is all positive and is unique up to the sign.

(2) *If u_h itself is the smallest eigenvector of $\mathcal{A}_{u_h,h}$, then it is also the unique global minimizer of $E_h(u)$.*

[4] An M-matrix is a matrix with nonnegative diagonal entries and nonpositive off-diagonal entries, with eigenvalues whose real parts are nonnegative.

For any other $w_h \neq \pm u_h$, we have

$$E_h(w_h) - E_h(u_h) = \|w_h\|^2_{\mathcal{A}_{u,h}} - \|u_h\|^2_{\mathcal{A}_{u,h}} + \frac{\beta}{2}\sum_{i=1}^{N}\left((w_h^{(i)})^4 + (u_h^{(i)})^4 - 2(w_h^{(i)})^2(u_h^{(i)})^2\right)h^d$$

$$= \left(\|w_h\|^2_{\mathcal{A}_{u,h}} - \|u_h\|^2_{\mathcal{A}_{u,h}}\right) + \frac{\beta}{2}\sum_{i=1}^{N}\left((w_h^{(i)})^2 - (u_h^{(i)})^2\right)^2 h^d > 0.$$

Thus u_h is the unique global minimizer of $E_h(u)$.

(3) *There is a unique positive eigenstate of (7.19), which is the ground state of (7.19) and the ground state of the linearized operator.*

Any positive iteration sequence stays positive with gradient descent iteration. The compactness of the constraint set ensures the existence of a sub-sequential limit point v_h, which is nonnegative. The fact that v_h is the minimizer of $E_h(u)$ implies that it is an eigenstate of $\mathcal{A}_{v,h}$. By Step (1), this eigenstate is all positive and is thus the smallest eigenstate of $\mathcal{A}_{v,h}$. By Step (2), it is also the unique global minimizer of $E_h(u)$.

□

Theorem 7.4.2 (Discrete version of Theorem 7.3.13)**.** *If the Sobolev PGD for $E_h(u)$ converges to the ground state v_h, and the step size $\{\tau_n\}$ satisfies $0 < \tau_{min} \leq \tau_n \leq \tau_{max}$, then it converges with an asymptotic exponential convergence rate.*

Proof. Theorem 7.4.1 ensures that v_h is still the "double" ground state of both $E_h(u)$ and $\mathcal{A}_{v_h,h}$. Thus, Theorems 7.3.9, 7.3.11, and 7.3.12 can all be generalized to the discretized case in the same way. The exponential convergence rate follows from the master theorem 7.2.1. □

Theorem 7.4.3 (Discrete version of Theorem 7.3.14)**.** *If the initial state u_0 satisfies $u_0(i) \geq 0\ \forall i$, and the step size $\{\tau_n\}$ satisfies $0 < \tau_{min} \leq \tau_n \leq \tau_{max}$, then the Sobolev PGD for $E_h(u)$ converges to the unique ground state v_h with an asymptotic exponential convergence rate.*

Proof. The proof follows similarly from the nonnegativity and uniqueness results of Theorem 7.4.1 and the exponential convergence result of Theorem 7.4.2. □

7.5 Generalization to other nonlinear eigenproblems

The Sobolev PGD points out a new direction for first-order fast solvers of nonlinear eigenproblems and higher (than quadratic) order optimization problems. Its application is thus well beyond the Gross-Pitaevskii eigenvalue problem. The operator class and the form of the objective function can be generalized. For example, consider

$$-\Delta v + Vv + \beta|v|^{2\alpha}v = \lambda v \tag{7.20}$$

for general $\alpha > 0$. This ground state equation and the corresponding time-dependent nonlinear Schrödinger equation are locally well-posed in $H^1(\mathbb{R}^d)$ as long as $2\alpha + 2 < \frac{2d}{\max\{d-2,0^+\}}$, see e.g., [58] and references therein. The previous Gross-Pitaevskii eigenvalue problem corresponds to the case $\alpha = 1$.

In general, Theorem 7.3.13 holds true for any $\alpha > 0$. The adaptive inner product remains well-posed and the ground state remains a "double" eigenstate. The change of inner product from $a_v(\cdot)$ to $a_u(\cdot)$ in the proof of Theorem 7.3.9 essentially relies on the convexity of the last term $\int |\cdot|^{2\alpha+2}$ in the energy functional $E(\cdot)$. Therefore, extensions of the previous results in both spatially continuous and discretized cases are easy. We do not present the details here.

It is also common in physics that the diffusion is not homogeneous in all spatial directions. For example, it can be stronger in two physical directions and weaker in the third one. More generally, we have

$$-\nabla \cdot (A(x)\nabla v) + Vv + \beta|v|^{2\alpha}v = \lambda v, \tag{7.21}$$

where the coefficient $A(x) \in L^\infty(\Omega)^{d\times d}$, $A(x)$ is symmetric and coercive. An interesting discrete counterpart to this is the nonlinear Schrödinger equation on metric trees (e.g. [50]), where the Laplacian is replaced by a graph Laplacian on a tree-graph $\mathcal{G}$.

When restricted to a bounded domain, so that the lowest part of the spectrum is always point spectrum, our previous arguments still hold. In the elliptic case, the discretized $\mathcal{A}_h$ may or may not be an M-matrix, but one can always turn to the small grid size limit $h \to 0^+$ limit when necessary.

For an even broader class of nonlinear eigenproblems or constrained optimization problems, the Sobolev gradient descent may still be applicable, but it is not clear whether *exponential convergence* is still true. It can be seen from previous sections that the convergence rate relies on the (L) condition, which in turn relies on the ground state v being the ground state of the linearized operator $\mathcal{A}_v$ at v, i.e., the so-called *"double ground state"* property. This is a nontrivial property in general, although it can be true for some operators like the biharmonic operator under certain conditions.

We discuss here one specific generalization of nonlinear Schrödinger eigenproblem, and demonstrate that the Sobolev PGD indeed has the potential of tackling previously formidable problems. The problem of interest is

$$-\Delta v + Vv + \beta|v|^2 v - \delta\Delta(|v|^2)v = \lambda v, \tag{7.22}$$

where $\delta \geq 0$. In other words, we add a higher-order interaction term $-\delta\Delta(|v|^2)$ to the Gross-Pitaevskii problem. The corresponding energy functional is

$$E(u) = \int |\nabla u|^2 + V|u|^2 + \frac{\beta}{2}|u|^4 + \frac{\delta}{2}\left|\nabla|u|^2\right|^2. \tag{7.23}$$

The above eigenproblem and its variational form are analyzed in [17]. Moreover, in [16] the authors propose to minimize the energy functional (7.23) by reformulating it as $E(\rho) = \int |\nabla\sqrt{\rho}|^2 + V\rho + \frac{\beta}{2}\rho^2 + \frac{\delta}{2}|\nabla\rho|^2$, where $\rho := |u|^2$. This reformulation facilitates the minimization, but it also suffers from the lack of continuity of $|\nabla\sqrt{\rho}|$ near $\rho \to 0^+$. This has to be treated with extra care, and a regularization term has to be added, which complicates the analysis. Therefore, instead of replacing $|u|^2$ with ρ, we propose to minimize $E(u)$ with respect to u directly with the Sobolev PGD.

Assume that Assumptions 7.3.1 still hold. Define the manifold $\mathcal{M}$ with an extra constraint:

$$\mathcal{M} := \left\{z \in H_0^1(\Omega) : \|u\|_{L^2} = 1,\ \|u\|_{L^\infty} \leq M_0,\ \|\nabla u\|_{L^\infty} \leq M_1\right\}.$$

Define the adaptive linearized operator and the respective inner products as follows:

$$\begin{aligned}
(z, w)_{a_u} &:= \int_\Omega \nabla z \nabla w + Vzw + \beta u^2 zw + \delta\nabla(uz)\nabla(uw), \\
(z, \mathcal{A}_u w)_{L^2} &:= (z, w)_{a_u}, \\
(z, w)_{a_0} &:= \int_\Omega \nabla z \nabla w + Vzw, \qquad \forall\, z,\ w \in \mathcal{T}_u\mathcal{M}, \quad u \in \mathcal{M}.
\end{aligned}$$

Then we have the following results.

Lemma 7.5.1. *The ground state v of (7.22) satisfies $v > 0$ everywhere on Ω. It is the unique positive eigenstate of (7.22). It is also both the unique ground state of (7.22) and that of the linearized operator $\mathcal{A}_v$ up to the sign.*

Proof. Following the same arguments as in Lemma 2 in [28], the extended $E(u)$ as in (7.23) still admits a nonnegative minimizer v. According to [17, Theorem 2.2], we know that $v \in C^{1,1}(\bar{\Omega})$. This implies that v, $\nabla v \in L^\infty(\Omega)$. Thus, the nonnegative v can still be made positive by the Harnack inequality. Also, the linearized operator $\mathcal{A}_v$ still has a unique positive ground state, which is exactly the above v. Thus the "double ground state" property remains true.

We now show that (7.22) has a unique positive eigenstate by a contradiction argument. Suppose instead that there is a different positive eigenstate $\tilde{v} > 0$ with its eigenvalue $\tilde{\lambda}$, and $E(\tilde{v}) > E(v)$. Using the Picone identity, $\int \nabla\tilde{v}\nabla(\frac{v^2}{\tilde{v}}) \leq \int (\nabla v)^2$. We have

$$\begin{aligned}
&\tilde{\lambda} - \lambda = \tilde{\lambda}(v,v)_{L^2} - (v,v)_{a_v} = \tilde{\lambda}\left(\tilde{v}, \frac{v^2}{\tilde{v}}\right)_{L^2} - (v,v)_{a_v} = \left(\tilde{v}, \frac{v^2}{\tilde{v}}\right)_{a_{\tilde{v}}} - (v,v)_{a_v} \\
&= \int \nabla\tilde{v}\cdot\nabla\left(\frac{v^2}{\tilde{v}}\right) + Vv^2 + \beta\tilde{v}^2v^2 + \delta\nabla(\tilde{v}^2)\nabla(v^2) - \int (\nabla v)^2 + Vv^2 + \beta v^4 + \delta(\nabla(v^2))^2 \\
&\leq \int (\nabla v)^2 + Vv^2 + \frac{\beta}{2}(v^4 + \tilde{v}^4) + \frac{\delta}{2}\left((\nabla(v^2))^2 + (\nabla(\tilde{v}^2))^2\right) - \int (\nabla v)^2 + Vv^2 + \beta v^4 + \delta(\nabla(v^2))^2 \\
&= \int \frac{\beta}{2}\tilde{v}^4 + \frac{\delta}{2}(\nabla(\tilde{v}^2))^2 - \int \frac{\beta}{2}v^4 + \frac{\delta}{2}(\nabla(v^2))^2 = (\tilde{\lambda} - E(\tilde{v})) - (\lambda - E(v)),
\end{aligned}$$

i.e.,

$$E(\tilde{v}) \leq E(v).$$

This contradicts our assumption that $E(\tilde{v}) > E(v)$. □

The next lemma shows that the eigenvalue and eigenfunction perturbation results stated in Lemma 7.3.7 hold similarly for (7.22).

Lemma 7.5.2. *Let λ_i and μ_i be the ith smallest eigenvalues of $\mathcal{A}_v$ and $\mathcal{A}_u$ respectively, and v_i and w_i be their corresponding eigenvectors (so that $v = v_1$). Let $C_v := \lambda_2 - \lambda_1$ denote the eigenvalue gap. Then there exists a positive constant $C = C(\beta, \delta, V, M_0, M_1, \Omega, \lambda_1, C_v)$, such that for all $\|u - v\|_{H^1} < C$, $u \in \mathcal{M}$, we have $\|u - w_1\|_{L^2} \leq s$ for some $s < 1$.*

Proof. The main idea of the proof is the same as that of Lemma 7.3.7 so we only point out their differences here. For example, to estimate $\mu_1 - \lambda_1$, we have

$$\begin{aligned}\mu_1 \le (v,v)_{a_u} &= (v,v)_{a_v} + \int_\Omega \beta(u^2v^2 - v^4) + \int_\Omega \delta\left((\nabla(uv)^2 - \nabla(v^2)^2)\right)\\ &= \lambda_1 + \int_\Omega \beta v^2(u+v)(u-v) + \int_\Omega \delta(\nabla(uv) + \nabla(v^2))(\nabla(uv) - \nabla(v^2)).\end{aligned}$$

The second term is bounded in the same way as the proof of Lemma 7.3.7. Only the third term containing high-order interaction is new. To bound the third term, we note that

$$\begin{aligned}&\int_\Omega \delta(\nabla(uv) + \nabla(v^2))(\nabla(uv) - \nabla(v^2))\\ &= \delta\int_\Omega (v\nabla u + u\nabla v + 2v\nabla v)(v\nabla u + u\nabla v - 2v\nabla v)\\ &\le 4\delta M_0 M_1 \int_\Omega |v\nabla u + u\nabla v - 2v\nabla v|\\ &= 4\delta M_0 M_1 \int_\Omega |v(\nabla u - \nabla v) + (u-v)\nabla v|\\ &\le C(\delta, M_0, M_1, \Omega)\|u - v\|_{H^1}.\end{aligned}$$

Similar bounds can be obtained in the estimation of $(\lambda_1 + \lambda_2) - (\mu_1 + \mu_2)$, $\lambda_1 - \mu_1$, and $(u,u)_{a_u} - \mu_1$. The dependence of the constant C only has two additional dependencies which are δ and M_1. □

Theorem 7.5.3. *If the initial state satisfies $u_0 \ge 0$ everywhere on Ω, then $\{u_n\}_{n=0}^\infty$ generated by the Sobolev PGD with step size $0 < \tau_{min} \le \tau_n \le \tau_{max}$ converges to the unique ground state v of (7.22) with an asymptotic exponential convergence rate.*

Proof. First, the Sobolev PGD sequence starting from a positive initial value remains positive as before, and convexity ensures convergence to a nonnegative local minimizer of $E(u)$, which must also be the global minimizer and the ground state of (7.22). This convergence can be proved to be a strong H^1 convergence by the Sobolev embedding and the convergence of energy.

In order to establish exponential convergence, it suffices to show that Conditions (L), (D), and (S) all hold for $\{u_n\}_{n=0}^\infty$. The nonnegativity of δ ensures the equivalence of a_0 and a_u norms. Thus Conditions (D) and (S) hold. Condition (L) follows from Lemma 7.5.2 and Lemma 7.3.8. □

The above results establish the exponential convergence of the Sobolev PGD for problem (7.22) for any $\delta \geq 0$. Numerical evidence shows that the Sobolev PGD for this problem converges very well just as the original Gross-Pitaevskii eigenproblem. This is a demonstration that the Sobolev gradient descent has the potential to be generalized to study some continuous or discrete high-degree optimization problems. We believe that this method has the potential to be extended to a broader class of problems as long as certain assumptions are satisfied, which is left for our future work.

7.6 Numerical experiments

In this section, we demonstrate the convergence of the Sobolev PGD method using some numerical examples. We show that exponential convergence rate is attained both for the original eigenproblem (7.1) and for its extension (7.22). We also observe and discuss some interesting phenomena that one may encounter in numerical experiments.

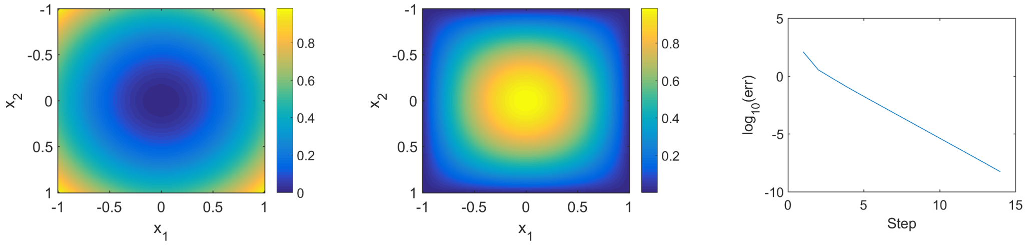
Gross-Pitaevskii eigenproblem in 2D.

We first look at the Gross-Pitaevskii eigenproblem (7.1) in two dimensions. Let the domain be $\Omega = [-1,1]^2 \subset \mathbb{R}^2$ with Dirichlet boundary condition. The problems are discretized with P1 Lagrange finite element method. The grid is a uniform grid with fixed size $h = 2 \cdot 2^{-8}$ throughout this section.

The first example is a single well potential $V(x) = \frac{1}{2}|x|^2$. It is well known that the Anderson localization [6] is present in this setting. We set $\beta = 1$. The initial guess z_0 is chosen as the eigenvector corresponding to the smallest eigenvalue of $\mathcal{A}_0$. It is strictly positive over the whole domain Ω. The step size is $\tau = 1$.

Figure 7.1a shows the profile of the potential $V(x)$. Figure 7.1b is the profile of the computed ground state with $\beta = 1$. Figure 7.1c displays the log H^1-error convergence $\log_{10}(\|u_n - v\|_{H^1}/\|v\|_{H^1})$. It can be seen that the Sobolev PGD converges in just a few steps with an exponential (linear) convergence rate.

By increasing β, there is a greater nonlinearity in the problem. When $\beta \gg 1$, the quartic term $\frac{\beta}{2}|u|^4$ would dominate the energy functional (7.2). This would be a significant barrier to some traditional methods. Yet the Sobolev PGD remains stable and fast. Figures 7.2a to 7.2d show the log H^1-error con-



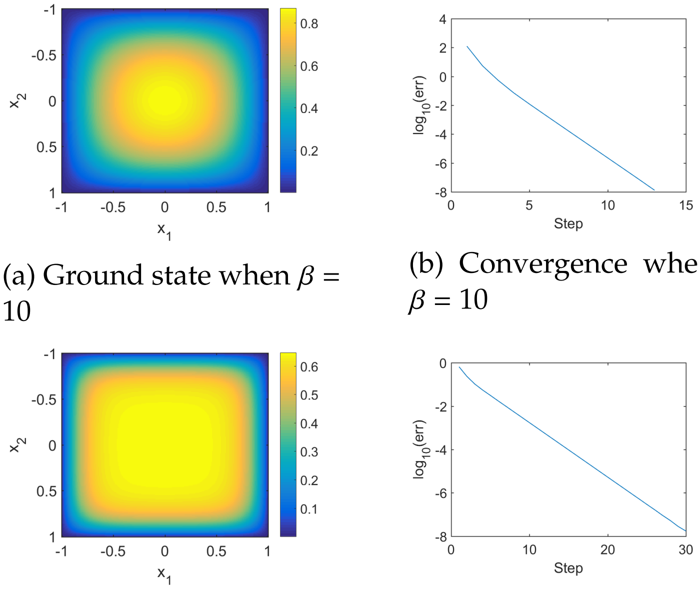
(a) Single well potential $V(x) = \frac{1}{2}|x|^2$

(b) Ground state when $\beta = 1$

(c) Log H^1-error convergence

Figure 7.1: Example of (7.1) with single well potential $V = \frac{1}{2}|x|^2$ and $\beta = 1$.

vergence and the profiles of the respective ground states with $\beta = 10$ and $\beta = 100$ respectively. With the Sobolev PGD, there is only a mild increase in the computational complexity, and the iteration still converges exponentially fast as predicted.



(a) Ground state when $\beta = 10$

(b) Convergence when $\beta = 10$

(c) Ground state when $\beta = 100$

(d) Convergence when $\beta = 100$

Figure 7.2: Example of (7.1) with single well potential $V = \frac{1}{2}|x|^2$ and $\beta = 10$ or 100.

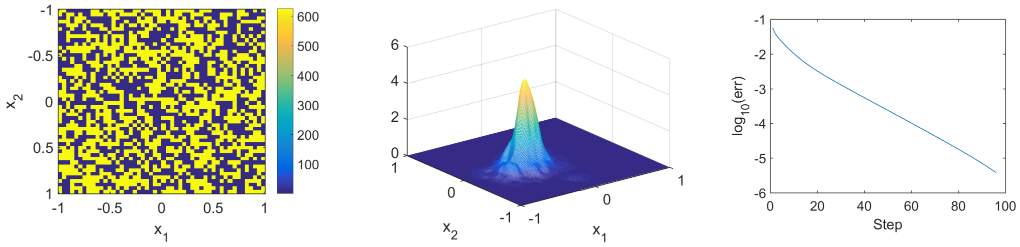
Localization under the disordered potential.

The second example is a disordered potential V. Its fully discrete counterpart, the randomized potential on the lattice $\mathbb{Z}^d$, has been extensively studied for its rich behaviour in spectral gaps, exponential localization of eigenstates near the bottom of the spectrum, and implications about the “mobility edge” conjecture in quantum physics and random matrix theory [48, 60].

In our semi-lattice example, the localization of the ground state is also present. In the experiment, $V(x)$ is generated as follows. The extent of disorder is determined by a parameter $K = 50$. This means that the domain Ω is divided into $K \times K$ cells. The value of V(x) in each cell is either 1 or $1/K^2$, randomly chosen with equal probability.

Figure 7.3a shows the profile of $V(x)$. Figure 7.3b displays the computed ground state with $\beta = 0.5$. It can be seen that the ground state is concentrated in a small region whose diameter is about a few times the interaction length of the disorder. Figure 7.3c shows the convergence rate of the Sobolev PGD iteration for this example.

To facilitate convergence, we have chosen $\tau = 1.5$. Although Theorem 7.3.14 requires a small τ, in the numerical experiments we find that choosing $\tau > 1$ results in significantly faster convergence. This is in accordance with the empirical findings of [72].



(a) Profile of the disordered potential

(b) Profile of the ground state

(c) Log H^1-error convergence

Figure 7.3: Example of (7.1) with a disordered potential and $\beta = 0.5$.

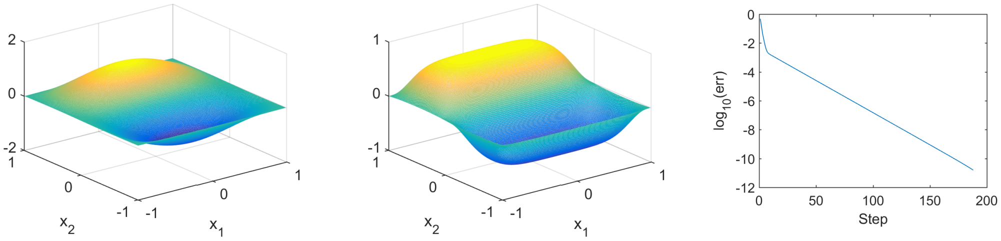
Asymptotic escape of Sobolev PGD from saddle states.

It is interesting to look at the asymptotic behaviour of the Sobolev gradient descent method if starting from a non-positive initial value. Recall that Theorem 7.3.14 only ensures exponential convergence to the global ground state from $u_0 \geq 0$. When this condition is violated, it is a priori unknown what the iteration will converge to. It is possible that there are other spurious fixed points, including local minimizers and saddle points. The first-order condition ensures that all these spurious fixed points are eigenstates. But the convergence rate to such points is unknown.

As for the spatially discretized case, the Hilbert manifold $\mathcal{M}$ becomes a Riemannian manifold, and the spectra of the operators become finite. As is proved in Chapter 3 and references therein, a random initialization almost surely avoids saddles and converges only to local or global minimizers. It means that if an excited state is a strict saddle point, then a random initialization is very unlikely to converge to that state. As for the spatially continuous case, it is reasonable to expect the same phenomenon, although the analysis could be more difficult due to the infinite dimension of $\mathcal{M}$ and the infinite number of eigenstates.

In the subsequent numerical experiments, we let $V(x) = \frac{1}{2}|x|^2$ and $\beta = 100$. We will use an example to show that for an excited state that is a strict saddle, it has a very thin converging set close to measure zero. Thus, using Sobolev PGD to compute excited states could be unstable. The accuracy of the computed excited states could be limited.

First, we let the initialization u_0 be the second-smallest eigenvector of $\mathcal{A}_0$. This u_0 is positive on half of Ω and negative on the other half. It is displayed in Figure 7.4a. We then let Sobolev PGD iterate a few steps. Figure 7.4b displays the computed state u^* when the algorithm stops. Figure 7.4c shows the decrease of the log L^2 error with respect to u^*. We also compute the manifold Hessian at u^* and find that it has at least one negative eigenvalue. Thus u^* is a strict saddle state.



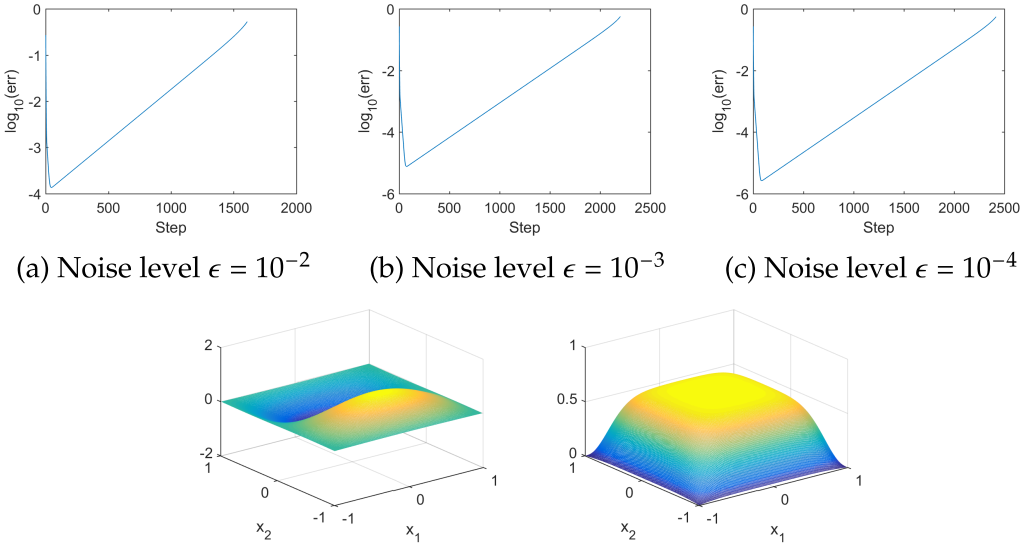
(a) Profile of initial state u_0

(b) Profile of computed state u^*

(c) Log H^1-error convergence

Figure 7.4: Behavior of Sobolev PGD for (7.1) with $u_0 \not\geq 0$.

Next, we add a small perturbation to u_0: we let $\hat{u}_0 = u_0 + \epsilon \cdot \eta$, where η is a random noise that is of the same order as u_0, and the parameter ϵ controls



(a) Noise level $\epsilon = 10^{-2}$ (b) Noise level $\epsilon = 10^{-3}$ (c) Noise level $\epsilon = 10^{-4}$

(d) Profile of $\hat{u}_0 = u_0 + 10^{-4}\eta$ (e) Profile of computed state starting from $\hat{u}_0$

Figure 7.5: Asymptotic escape from saddle state under small perturbations. Figures (a)–(c) displays the distances to the saddle state u^* starting from $\hat{u}_0 = u_0 + \epsilon \cdot \eta$.

the magnitude of noise. We let Sobolev PGD start from $\hat{u}_0$ and trace its evolution. What we observe is that, as long as there is a small perturbation, Sobolev PGD escapes from the previous saddle state and converges to the ground state. The parameter ϵ can be chosen as small as 10^{-4} and this effect is still present.

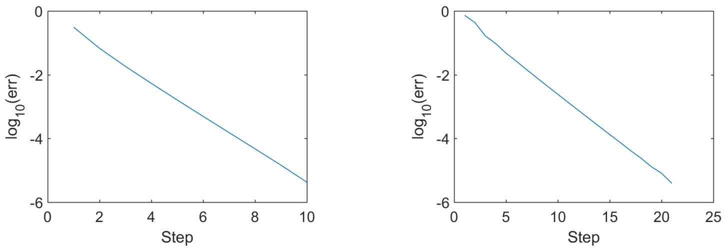
Specifically, Figures 7.5a to 7.5c demonstrate the evolution of the log-distance to the precomputed closest excited state u^*. We choose $\epsilon = 10^{-2}$, 10^{-3}, and 10^{-4}, respectively. Saddle escape behavior can be observed in all three cases. We can see that the distance to the excited state first goes down, then goes up. Figure 7.5e show the computed state starting from $\hat{u}_0$, and it is the ground state.

In general, first-order optimization methods, including Sobolev PGD as well as other methods in the gradient descent family, are not good choices for the computation of excited states. They rely on a good enough initialization (like the above u_0 without noise) and could suffer from numerical instability issues. One has to resort to other methods if the goal is to obtain high accuracy. This would be an interesting topic for future research.

High-order interaction.

We now look at Problem (7.22) with an extra high-order interaction term. This adds additional nonlinearity to the problem. Consider the same domain $\Omega = [-1, 1]^2 \subset \mathbb{R}^2$ and spatial discretization size $h = 2 \cdot 2^{-8}$. Let $V(x) = \frac{1}{2}|x|^2$ still be the single well potential. The first example is $\beta = 10$ and $\delta = 1$. Figure 7.6a shows the log error convergence. The iteration converges in a few steps and shows a good convergence rate.

In the second example, we increase δ and look at the problem with strong high-order interaction. We choose $\beta = 100$ and $\delta = 100$. Figure 7.6b shows the log error convergence. The convergence rate is slower but stable.



(a) Convergence when $\beta = 10$, $\delta = 1$

(b) Convergence when $\beta = 100$, $\delta = 100$

Figure 7.6: Examples of (7.22) with different nonlinear effects

7.7 Discussion

In this chapter, we analyzed the exponential convergence of the a_u-Sobolev gradient descent method without resorting to the time-continuous gradient flow. To this purpose, we introduced a general convergence tool using the Łojasiewicz inequality, and adapted it to the setting of infinite dimensional Hilbert manifold and mixed norms. By proving the (L), (D), and (S) conditions for the Sobolev PGD, we were able to unveil the mechanism behind the good performance of the Sobolev PGD for the Gross-Pitaevskii eigenproblem (7.1), which was only empirically observed in previous works.

The success of the Sobolev PGD on the Gross-Pitaevskii eigenproblem inspires us to further explore alternative fast solvers for more general nonlinear eigenproblems and optimizations with high-degree objective functions. Our analysis revealed that the essential condition is the “double ground state” property, namely the ground state of the nonlinear problem is also the unique ground state of the linearized operator at that point. This can be rigorously proved in some cases and seems to be true in a number of physical applications of interest based on empirical evidence. Specifically, we showed that this condition is satisfied for a nonlinear Schrödinger eigenproblem with extra high-order interaction term. Thus the Sobolev PGD works well for this problem and has superiority over previous methods.

Bibliography

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] P-A Absil, Robert Mahony, and Jochen Trumpf. An extrinsic look at the Riemannian Hessian. In *International Conference on Geometric Science of Information*, pages 361–368. Springer, 2013.

[3] Pierre-Antoine Absil, Robert Mahony, and Benjamin Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization, 16(2):531–547*, 2005.

[4] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[5] Mike H Anderson, Jason R Ensher, Michael R Matthews, Carl E Wieman, and Eric A Cornell. Observation of Bose-Einstein condensation in a dilute atomic vapor. *Science*, pages 198–201, 1995.

[6] Philip W Anderson. Absence of diffusion in certain random lattices. *Physical review*, 109(5):1492, 1958.

[7] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[8] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

[9] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[10] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[11] Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. *arXiv preprint arXiv:1810.10702*, 2018.

[12] Augustin Banyaga and David Hurtubise. Morse-Bott homology. *Transactions of the American Mathematical Society*, 362(8):3997–4043, 2010.

[13] Augustin Banyaga and David Hurtubise. *Lectures on Morse homology*, volume 29. Springer Science & Business Media, 2013.

[14] Weizhu Bao and Yongyong Cai. Mathematical theory and numerical methods for Bose-Einstein condensation. *Kinetic & Related Models*, 6 (1):1–135, 2013.

[15] Weizhu Bao and Qiang Du. Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow. *SIAM Journal on Scientific Computing*, 25(5):1674–1697, 2004.

[16] Weizhu Bao and Xinran Ruan. Computing ground states of Bose-Einstein condensates with higher order interaction via a regularized density function formulation. *arXiv preprint arXiv:1908.09096*, 2019.

[17] Weizhu Bao, Yongyong Cai, and Xinran Ruan. Ground states of Bose–Einstein condensates with higher order interaction. *Physica D: Nonlinear Phenomena*, 386:38–48, 2019.

[18] James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, 2007.

[19] Andrea L Bertozzi and Arjuna Flenner. Diffuse interface models on graphs for classification of high dimensional data. *Multiscale Modeling & Simulation*, 10(3):1090–1118, 2012.

[20] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*, 29, 2016.

[21] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6): 3319–3363, 2010.

[22] Angelika Bunse-Gerstner, Ralph Byers, Volker Mehrmann, and Nancy K Nichols. Numerical computation of an analytic singular value decomposition of a matrix valued function. *Numerische Mathematik*, 60(1):1–39, 1991.

[23] HanQin Cai, Jian-Feng Cai, and Ke Wei. Accelerated alternating projections for robust principal component analysis. *The Journal of Machine Learning Research*, 20(1):685–717, 2019.

[24] Jian-Feng Cai and Ke Wei. Solving systems of phaseless equations via Riemannian optimization with optimal sampling complexity. *arXiv preprint arXiv:1809.02773*, 2018.

[25] Jian-Feng Cai, Hui Ji, Zuowei Shen, and Gui-Bo Ye. Data-driven tight frame construction and image denoising. *Applied and Computational Harmonic Analysis*, 37(1):89–105, 2014.

[26] Léopold Cambier and P-A Absil. Robust low-rank matrix completion by Riemannian optimization. *SIAM Journal on Scientific Computing*, 38 (5):S440–S460, 2016.

[27] Eric Cancès and Claude Le Bris. On the convergence of SCF algorithms for the Hartree-Fock equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 34(4):749–774, 2000.

[28] Eric Cancès, Rachida Chakir, and Yvon Maday. Numerical analysis of nonlinear eigenvalue problems. *Journal of Scientific Computing*, 45 (1-3):90–117, 2010.

[29] Eric Cancès, Gaspard Kemlin, and Antoine Levitt. Convergence analysis of direct minimization and self-consistent iterations. *arXiv preprint arXiv:2004.09088*, 2020.

[30] Emmanuel J Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[31] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9 (6):717, 2009.

[32] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[33] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[34] Emmanuel J Candès, 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.

[35] Volkan Cevher, Aswin Sankaranarayanan, Marco F Duarte, Dikpal Reddy, Richard G Baraniuk, and Rama Chellappa. Compressive sensing for background subtraction. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10*, pages 155–168. Springer, 2008.

[36] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[37] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Constantine Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013.

[38] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37, 2019.

[39] 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.

[40] Ralph Cohen. *Topics in Morse theory*. Stanford University Department of Mathematics, 1991.

[41] Christopher Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. In *Advances in Neural Information Processing Systems*, pages 5985–5995, 2019.

[42] Ionut Danaila and Parimah Kazemi. A new Sobolev gradient method for direct minimization of the Gross–Pitaevskii energy with rotation. *SIAM Journal on Scientific Computing*, 32(5):2447–2467, 2010.

[43] Ionut Danaila and Bartosz Protas. Computation of ground states of the Gross–Pitaevskii functional via Riemannian optimization. *SIAM Journal on Scientific Computing*, 39(6):B1102–B1129, 2017.

[44] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7 (1):1–46, 1970.

[45] Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018.

[46] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, 2020.

[47] Christopher M De Sa, Satyen Kale, Jason D Lee, Ayush Sekhari, and Karthik Sridharan. From gradient flow on population loss to learning with stochastic gradient descent. *Advances in Neural Information Processing Systems*, 35:30963–30976, 2022.

[48] Percy Deift et al. Some open problems in random matrix theory and the theory of integrable systems. II. *SIGMA. Symmetry, Integrability and Geometry: Methods and Applications*, 13:016, 2017.

[49] Luca Dieci and Timo Eirola. On smooth decompositions of matrices. *SIAM Journal on Matrix Analysis and Applications*, 20(3):800–819, 1999.

[50] Simone Dovetta, Enrico Serra, and Paolo Tilli. NLS ground states on metric trees: existence results and open questions. *arXiv preprint arXiv:1905.00655*, 2019.

[51] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in neural information processing systems*, pages 1067–1077, 2017.

[52] Geneviève Dusson and Yvon Maday. A posteriori analysis of a nonlinear Gross–Pitaevskii-type eigenvalue problem. *IMA Journal of Numerical Analysis*, 37(1):94–137, 2017.

[53] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[54] Erwan Faou and Tiphaine Jézéquel. Convergence of a normalized gradient algorithm for computing ground states. *IMA Journal of Numerical Analysis*, 38(1):360–376, 2018.

[55] Salar Fattahi and Somayeh Sojoudi. Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis. *Journal of machine learning research*, 2020.

[56] James R Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

[57] Catherine Fraikin, K Hüper, and P Van Dooren. Optimization over the Stiefel manifold. In *PAMM: Proceedings in Applied Mathematics and Mechanics*, volume 7, pages 1062205–1062206. Wiley Online Library, 2007.

[58] Rupert L Frank. Ground states of semi-linear PDEs. *Lecture notes*, 2014.

[59] Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.

[60] Jürg Fröhlich and Thomas Spencer. Absence of diffusion in the Anderson tight binding model for large disorder or low energy. *Communications in Mathematical Physics*, 88(2):151–184, 1983.

[61] Komei Fukuda. Lecture: Polyhedral computation, Spring 2015. *Department of Mathematics and Institute of Theoretical Computer Science, ETH Zurich, Switzerland*, 2015.

[62] CB Garcia and Tien-Yian Li. On the number of solutions to polynomial systems of equations. *SIAM Journal on Numerical Analysis*, 17(4): 540–546, 1980.

[63] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

[64] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

[65] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

[66] Ralph W Gerchberg. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[67] David Gilbarg and Neil S Trudinger. *Elliptic partial differential equations of second order*. springer, 2015.

[68] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[69] Jun He, Laura Balzano, and John Lui. Online robust subspace tracking from partial information. *arXiv preprint arXiv:1109.3827*, 2011.

[70] Uwe Helmke and John B Moore. *Optimization and dynamical systems*. Springer Science & Business Media, 2012.

[71] Uwe Helmke and Mark A Shayman. Critical points of matrix least squares distance functions. *Linear Algebra and its Applications*, 215:1–19, 1995.

[72] Patrick Henning and Daniel Peterseim. Sobolev gradient flow for the Gross-Pitaevskii eigenvalue problem: Global convergence and computational efficiency. *SIAM Journal on Numerical Analysis*, 58(3):1744–1772, 2020.

[73] Patrick Henning, Axel Målqvist, and Daniel Peterseim. Two-level discretization techniques for ground state computations of Bose-Einstein condensates. *SIAM Journal on Numerical Analysis*, 52(4):1525–1550, 2014.

[74] Sebastian Holtz, Thorsten Rohwedder, and Reinhold Schneider. On manifolds of tensors of fixed TT-rank. *Numerische Mathematik*, 120(4): 701–731, 2012.

[75] Thomas Y Hou, De Huang, Ka Chun Lam, and Ziyun Zhang. A fast hierarchically preconditioned eigensolver based on multiresolution matrix decomposition. *Multiscale Modeling & Simulation*, 17(1):260–306, 2019.

[76] Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. Analysis of asymptotic escape of strict saddle sets in manifold optimization. *SIAM Journal on Mathematics of Data Science*, 2(3):840–871, 2020.

[77] Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. Fast global convergence for low-rank matrix recovery via Riemannian gradient descent with random initialization. *arXiv preprint arXiv:2012.15467*, 2020.

[78] Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. Asymptotic escape of spurious critical points on the low-rank matrix manifold. *arXiv preprint arXiv:2107.09207*, 2021.

[79] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[80] Jiang Hu, Xin Liu, Zai-Wen Wen, and Ya-Xiang Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020.

[81] Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi. Phase retrieval: An overview of recent developments. *arXiv preprint arXiv:1510.07713*, 2015.

[82] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

[83] Tetsuya Kaneko, Simone Fiori, and Toshihisa Tanaka. Empirical arithmetic averaging over the compact Stiefel manifold. *IEEE Transactions on Signal Processing*, 61(4):883–894, 2012.

[84] Fatih Kangal, Karl Meerbergen, Emre Mengi, and Wim Michiels. A subspace method for large-scale eigenvalue optimization. *SIAM Journal on Matrix Analysis and Applications*, 39(1):48–82, 2018.

[85] Othmar Koch and Christian Lubich. Dynamical low-rank approximation. *SIAM Journal on Matrix Analysis and Applications*, 29(2):434–454, 2007.

[86] Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014.

[87] R de L Kronig and William George Penney. Quantum mechanics of electrons in crystal lattices. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 130 (814):499–513, 1931.

[88] James R Kuttler. Finite difference approximations for eigenvalues of uniformly elliptic operators. *SIAM Journal on Numerical Analysis*, 7(2): 206–232, 1970.

[89] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.

[90] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1): 311–337, 2019.

[91] John Lee. *Introduction to Smooth Manifolds*, volume 218. Springer Science & Business Media, 2012.

[92] Eitan Levin, Joe Kileel, and Nicolas Boumal. Finding stationary points on bounded-rank matrices: A geometric hurdle and a smooth remedy. *Mathematical Programming*, pages 1–34, 2022.

[93] Tien-Yien Li. Solving polynomial systems. *The mathematical intelligencer*, 9(3):33–39, 1987.

[94] Wuchen Li and Guido Montúfar. Natural gradient via optimal transport. *Information Geometry*, 1(2):181–214, 2018.

[95] Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *arXiv preprint arXiv:1911.05047*, 2019.

[96] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020.

[97] Xinrong Li, Naihua Xiu, and Ziyan Luo. Low-rank matrix optimization over affine set. *arXiv preprint arXiv:1912.03029*, 2019.

[98] Zhenzhen Li, Jian-Feng Cai, and Ke Wei. Toward the optimal construction of a loss function without spurious local minima for solving quadratic equations. *IEEE Transactions on Information Theory*, 66(5): 3242–3260, 2019.

[99] Elliott H Lieb, Robert Seiringer, and Jakob Yngvason. Bosons in a trap: A rigorous derivation of the Gross-Pitaevskii energy functional. In *The Stability of Matter: From Atoms to Stars*, pages 685–697. Springer, 2001.

[100] Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117: 87–89, 1963.

[101] Stanislaw Lojasiewicz. Ensembles semi-analytiques, preprint 112 pp. *IHES notes. Available at http://perso.univ-rennes1.fr/michel.coste/Lojasiewicz.pdf.*, 1965.

[102] Christian Lubich, Ivan V Oseledets, and Bart Vandereycken. Time integration of tensor trains. *SIAM Journal on Numerical Analysis*, 53(2): 917–941, 2015.

[103] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

[104] Jan R Magnus. On differentiating eigenvalues and eigenvectors. *Econometric Theory*, pages 179–191, 1985.

[105] Estelle Massart and P-A Absil. Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, 41(1):171–198, 2020.

[106] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[107] Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108 (1):177–205, 2006.

[108] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust PCA. *arXiv preprint arXiv:1410.7660*, 2014.

[109] Guillaume Olikier, Kyle A Gallivan, and P-A Absil. An apocalypse-free first-order low-rank optimization algorithm. *arXiv preprint arXiv:2201.03962*, 2022.

[110] Donal B O'Shea and Leslie C Wilson. Limits of tangent spaces to real surfaces. *American journal of mathematics*, 126(5):951–980, 2004.

[111] Vidvuds Ozoliņš, Rongjie Lai, Russel Caflisch, and Stanley Osher. Compressed modes for variational problems in mathematics and physics. *Proceedings of the National Academy of Sciences*, 110(46):18368–18373, 2013.

[112] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*, 2016.

[113] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2233–2246, 2012.

[114] Lawrence Perko. *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media, 2013.

[115] Lev Pitaevskii and Sandro Stringari. *Bose-Einstein condensation and superfluidity*, volume 164. Oxford University Press, 2016.

[116] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[117] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[118] Reinhold Schneider and André Uschmajew. Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality. *SIAM Journal on Optimization*, 25(1):622–646, 2015.

[119] 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.

[120] Michael Shub. *Global stability of dynamical systems*. Springer Science & Business Media, 2013.

[121] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

[122] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18:1131–1198, 2018.

[123] Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 7274–7284, 2019.

[124] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[125] Bart Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

[126] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

[127] Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

[128] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

[129] Ming Yan, Yi Yang, and Stanley Osher. Exact low-rank matrix completion from sparsely corrupted entries via adaptive outlier pursuit. *Journal of Scientific Computing*, 56:433–449, 2013.

[130] Ke Ye, Ken Sze-Wai Wong, and Lek-Heng Lim. Optimization on flag manifolds. *arXiv preprint arXiv:1907.00949*, 2019.

[131] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

[132] Haixiang Zhang, Andre Milzarek, Zaiwen Wen, and Wotao Yin. On the geometric analysis of a quartic-quadratic optimization problem under a spherical constraint. *arXiv preprint arXiv:1908.00745*, 2019.

[133] Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *Journal of Machine Learning Research*, 20 (114):1–34, 2019.

[134] Ziyun Zhang. Exponential convergence of Sobolev gradient descent for a class of nonlinear eigenproblems. *Communications in Mathematical Sciences*, 20(2):377–403, 2022.

[135] Zhihui Zhu, Yifan Wang, Daniel Robinson, Daniel Naiman, Rene Vidal, Manolis C Tsakiris, et al. Dual principal component pursuit: Improved analysis and efficient algorithms. In *Neural Information Processing Systems (NIPS)*, 2018.