When Are Nonconvex Optimization Problems Not Scary?

Ju Sun

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

When Are Nonconvex Optimization Problems Not Scary?

Ju Sun

Nonconvex optimization is NP-hard, even the goal is to compute a local minimizer. In applied disciplines, however, nonconvex problems abound, and simple algorithms, such as gradient descent and alternating direction, are often surprisingly effective. The ability of simple algorithms to find high-quality solutions for practical nonconvex problems remains largely mysterious.

This thesis focuses on a class of nonconvex optimization problems which *can* be solved to global optimality with polynomial-time algorithms. This class covers natural nonconvex formulations of central problems in signal processing, machine learning, and statistical estimation, such as sparse dictionary learning (DL), generalized phase retrieval (GPR), and orthogonal tensor decomposition. For each of the listed problems, the nonconvex formulation and optimization lead to novel and often improved computational guarantees.

This class of nonconvex problems has two distinctive features: (i) *All local minimizer are also global.* Thus obtaining any local minimizer solves the optimization problem; (ii) *Around each saddle point or local maximizer, the function has a negative directional curvature.* In other words, around these points, the Hessian matrices have negative eigenvalues. We call smooth functions with these two properties (qualitative) $\mathcal{X}$ functions, and derive concrete quantities and strategy to help verify the properties, particularly for functions with random inputs or parameters. As practical examples, we establish that certain natural nonconvex formulations for complete DL and GPR are $\mathcal{X}$ functions with concrete parameters.

Optimizing $\mathcal{X}$ functions amounts to finding any local minimizer. With generic initializations, typical iterative methods at best only guarantee to converge to a critical point that might be a saddle point or local maximizer. Interestingly, the $\mathcal{X}$ structure allows a number of iterative methods to escape from saddle points and local maximizers and efficiently find a local minimizer, without special initializations. We choose to describe and analyze the second-order trust-region method (TRM) that seems to yield the strongest computational guarantees. Intuitively, second-order methods can exploit Hessian to extract negative curvature directions

around saddle points and local maximizers, and hence are able to successfully escape from the saddles and local maximizers of $\mathcal{X}$ functions. We state the TRM in a Riemannian optimization framework to cater to practical manifold-constrained problems. For DL and GPR, we show that under technical conditions, the TRM algorithm finds a global minimizer in a polynomial number of steps, from arbitrary initializations.

Table of Contents

List of Figures

Notations

$\mathbb{R}^n$	n-dimensional real space
$\mathbb{C}^n$	n-dimensional complex space
$\mathbb{B}^n$, $\mathbb{CB}^n$	unit ball in $\mathbb{R}^n$, $\mathbb{C}^n$
$\mathbb{S}^{n-1}$, $\mathbb{CS}^{n-1}$	unit sphere in $\mathbb{R}^n$, $\mathbb{C}^n$
$\Re(\boldsymbol{z})$, $\Im(\boldsymbol{z})$	real, complex parts (as vectors) of a complex vector $\boldsymbol{z}$
O_n	orthogonal group of order n
$\boldsymbol{X}$	bold capital letters as matrices
$\boldsymbol{x}$	bold small letters as vectors
$\boldsymbol{x}^i$	i-th row of $\boldsymbol{X}$ as column vector
$\boldsymbol{x}_j$	j-th column of $\boldsymbol{X}$ as column vector
$\lVert\cdot\rVert$	vector ℓ^2 norm or matrix operator norm
$\lVert\cdot\rVert_F$	matrix Frobenius norm
$(\cdot)^\top$	transposition without conjugation
$(\cdot)^*$	conjugate transposition, equivalent to $(\cdot)^\top$ for real vectors/matrices; preferred over $(\cdot)^\top$ when no confusion caused
$\doteq$	defined as
$[k]$	the set $\{1, \dots, k\}$
C, c, C_k, c_k	C, c and all indexed versions for absolute constants with **local scopes**
$X \sim \mathcal{L}$	random variable X distributed by the law $\mathcal{L}$
$\mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$	standard Gaussian distribution in $\mathbb{R}^n$
$\mathcal{CN}(n)$	standard complex Gaussian distribution in $\mathbb{C}^n$
$\mathrm{Ber}(\theta)$	standard Bernoulli distribution with parameter θ
$X \sim_{i.i.d.} \mathcal{L}$	elements in (vector- or matrix-valued) X independent, identically distributed by the law $\mathcal{L}$

$X \sim \mathrm{BG}(\theta)$	$X = W \cdot Z$ with $W \sim \mathrm{Ber}(\theta)$ and independently $W \sim \mathcal{N}(0, 1)$
w.h.p.	short for “with high probability”
i.i.d.	short for “independent, identically distributed”
w.l.o.g.	short for “without loss of generality”
w.r.t.	short for “with respect to”
(C)DL	short for “(complete) dictionary learning”
(C)DR	short for “(complete) dictionary recovery”
GPR	short for “generalized phase retrieval”
TRM	short for “trust region method”

Acknowledgments

I would like to give my foremost thanks to my advisor Professor John Wright. His high intellectual standard and enthusiasm for research have transformed entirely the path of my studies and research. His hand-on advising style and great patience on working out details with students have largely mitigated the pain to enter theoretical research. His wealth of ideas, clarity of thought, insistence on solving fundamental problems, and sense of humor have made my PhD journey exceptionally enriching and enjoyable. I am also eternally indebted to Professor Loong-Fah Cheong and Professor Shuicheng Yan at the National University of Singapore, who set me on the road. Professor Cheong's course on 3D vision piqued my interest in computer vision and applied mathematics, and Professor Yan brought me to the forefront of visual recognition and machine learning in the first few months of my post-graduate research.

I am especially grateful to Professor Shih-Fu Chang, Professor Weinan E (Princeton U.), Professor Donald Goldfarb, Professor Xiaodong Wang for serving on my thesis committee. Professor Chang discussed in detail with me the dictionary learning work after my proposal, which has greatly helped me improve the relevant presentation. Professor E (and his students) had two long discussions with me on the $\mathcal{X}$ structure that underpins the thesis. His numerous insightful questions and comments have set me to rethink many parts of the works/presentation, and added more open questions to my mental list. In addition, I would like to thank those who have hosted talks for me on research covered in this thesis: Dr. Alekh Agarwal (Microsoft NYC), Professor Emmanuel Candès (Stanford U.), Professor Qiang Du (Columbia U.), Professor Gilad Lerman (Minnesota U.), Professor Amit Singer (Princeton U.).

The Wei Family Private Foundation selected me to be among its three inaugural fellows and paid the tuition for my first three years. I would like to express my deepest respect and gratitude to the deceased Mr. and Mrs. Wei, and their good friends who are running the foundation, for their unconditional support! The fellowship is an invaluable resource for members in a start-up group.

I am fortunate to have had great colleagues and friends throughout the five years. I would like to thank my fellow group members, Yuqian Zhang, Henry Kuo, Qing Qu, Yenson Lau, Cun Mu, Zhengyu Chen, and Robert Colgan, for making the group a vibrant and friendly one. Among them, I especially thank Qing Qu, my close collaborator, who should be accorded credit for significant part of this thesis. Our collaboration has moved things forward at a pace that I would never anticipate. At Columbia and about, there are numerous other friends that I am afraid I cannot list exhaustively here – I owe all of them a big thanks! I would like to especially acknowledge two groups of friends, organized grossly into two (nonpolitical!) parties: the hotpot party and the wine (not tea!) party. You know, delicious food and wine are indispensable, especially when research life runs dry! My thanks also extend outside Columbia to many other friends: Nicolas Boumal (Princeton U.), Yuxin Chen (Stanford U.), Ruoyu Sun (Stanford U.), Zhaoran Wang (Princeton U.), Veniamin Morgenshtern (Stanford U.) have recently discussed my research problems with me. My thanks also go across the Pacific to a tiny tropical country called Singapore, where most of my undergrad friends are now inhabiting, and of course also to China, my motherland!

Thesis acknowledgement usually does not include pre-college teachers, which I think is unfair. I seriously think so because I grew up in circumstances where education resources were severely limited: in rural areas of China – a great teacher means everything! My special thanks go to all the teachers I met in my pre-college life! Out of the many great teachers I encountered, I would like to gratefully acknowledge three, coincidentally all teaching mathematics: Mr. Wencai Fu, who taught me primary-school math, let me know the integral sign and the intuition about infinitesimal; Ms. Guiqin Li, who taught me secondary-school math, assured me self-learning is possible and more effective; Mr. Tao Yang, who taught me high-school math, brought me out of waning period of life via care and trust. Without them, I would never get the chance to embark on a research journey.

Last but not least, my heartfelt gratitude goes to my family, particularly my parents! My parents did not even get chance to receive secondary-school education. But to me they are undoubtedly the most wonderful parents: they grant me invaluable freedom that bears their unconditional love, trust, and support!

Ju Sun
May 5, 2016
New York

To my parents

Part I

Overview

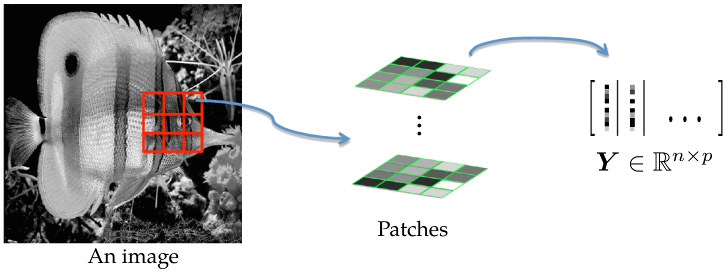
Chapter 1

Introduction

> Everything should be made as simple as possible, but no simpler.
>
> Albert Einstein

The whole line of research contained in this thesis was inspired by a curious experiment, which concerns learning compact representation for a given data collection.

1.1 An intriguing experiment with real images



We focus on learning compact representation for a collection of image patches. Specifically, we divide a given greyscaled image into non-overlapping patches, which are then converted into vectors and stacked column-wise into a data matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$. The task is seeking a factorization $\boldsymbol{A}\boldsymbol{X}$ such that:

$$\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X},\ \boldsymbol{A} \in \mathbb{R}^{n\times m}, \boldsymbol{X} \in \mathbb{R}^{m\times p}, \quad \text{and} \quad \boldsymbol{X} \text{ as sparse as possible.}$$

Here one can think of columns of $\boldsymbol{A}$ as a representation basis for the given image patches, and $\boldsymbol{X}$ as the coefficients that encode the patches with respect to the basis. Finding such sparsifying basis for given visual data proves critical to image compression and classification [DeV98, DVDD98, Tem03, DeV09, Can02, MP10a, Ela10, MBP14].

For simplicity, we shall try to learn orthogonal $\boldsymbol{A}$, and consider a natural nonconvex formulation

$$\mathrm{minimize}_{\boldsymbol{A}\in\mathbb{R}^{n\times n},\boldsymbol{X}\in\mathbb{R}^{n\times p}}\ \lambda\left\|\boldsymbol{X}\right\|_1+\frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{X}-\boldsymbol{Y}\right\|_F^2,\ \text{subject to}\ \ \boldsymbol{A}\in O_n. \tag{1.1.1}$$

Formulation (1.1.1) attempts to find a pair $(\boldsymbol{A},\boldsymbol{X})$ that best trades off sparsity and fidelity to the observed data $\boldsymbol{Y}$. Here $\left\|\cdot\right\|_1\doteq\sum_{i,j}|X_{ij}|$ promotes sparsity of the coefficients $\boldsymbol{X}$, $\left\|\cdot\right\|_F$ is the usual Frobenius norm of matrices, and λ is a tunable parameter that trades off coefficient sparsity and quality of approximation. O_n denotes the set of orthogonal matrices in $\mathbb{R}^{n\times n}$, i.e., orthogonal group of order n.

Problem (1.1.1) is in no way convex. The objective is nonconvex due to the bilinear map $(\boldsymbol{A},\boldsymbol{X})\mapsto\boldsymbol{A}\boldsymbol{X}$. More interestingly, this bilinear map induces intrinsic symmetry to the optimization space. Indeed, for any pair of feasible $(\boldsymbol{A},\boldsymbol{X})$, $(\boldsymbol{A}\boldsymbol{\Pi}\boldsymbol{\Sigma},\boldsymbol{\Sigma}^{-1}\boldsymbol{\Pi}^{-1}\boldsymbol{X})$ for all permutation matrix $\boldsymbol{\Pi}$ and all diagonal sign matrix[1] $\boldsymbol{\Sigma}$ produce exactly the same objective value to (1.1.1). This implies that there are combinatorially many global minimizers to (1.1.1), and they are generally[2] isolated from each other in the space! Moreover, the orthogonal group O_n is a nonconvex set.

To derive a concrete algorithm for (1.1.1), one can deploy the alternating direction method (ADM)[3], i.e., alternately minimizing the objective function with respect to one variable while fixing the other. The iteration sequence actually takes a very simple form: for $k=1,2,3,\dots$,

$$\boldsymbol{X}_k=\mathcal{S}_\lambda\left[\boldsymbol{A}_{k-1}^\top\boldsymbol{Y}\right],\qquad \boldsymbol{A}_k=\boldsymbol{U}\boldsymbol{V}^\top\ \text{for}\ \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top=\mathtt{SVD}\left(\boldsymbol{Y}\boldsymbol{X}_k^\top\right)^{4}$$

where $\mathcal{S}_\lambda\left[\cdot\right]$ denotes the well-known soft-thresholding operator [DJ94, Don95] acting elementwise on matrices, i.e., $\mathcal{S}_\lambda\left[x\right]\doteq\mathrm{sign}\left(x\right)\max\left(|x|-\lambda,0\right)$ for any scalar x.

Figure 1.1 shows what we obtained using the simple ADM algorithm, with *independent and randomized initializations*:

Across many natural images, the algorithm seems to always produce the same solution for each

[1] ...i.e., with ±1 on the diagonal.

[2] For a fixed global minimizer $(\boldsymbol{A}_\star,\boldsymbol{X}_\star)$, it is easy to see that all its equivalent copies due to signed permutations are isolated from each other. The situation gets slightly complicated when there are other global minimizers that are not equivalent to $(\boldsymbol{A}_\star,\boldsymbol{X}_\star)$ by the intrinsic symmetry. Our comment pertains to the generic case when the set of all the global minimizers does not form a connected subset in the product space $O_n\times\mathbb{R}^{n\times p}$.

[3] This method is also called alternating minimization or (block) coordinate descent method. see, e.g., [BT89, Tse01] for classic results and [ABRS10, BST14] for several interesting recent developments.

[4] In other words, $\boldsymbol{A}_k$ is the orthogonal matrix arising from the polar decomposition of $\boldsymbol{Y}\boldsymbol{X}_k^\top$.

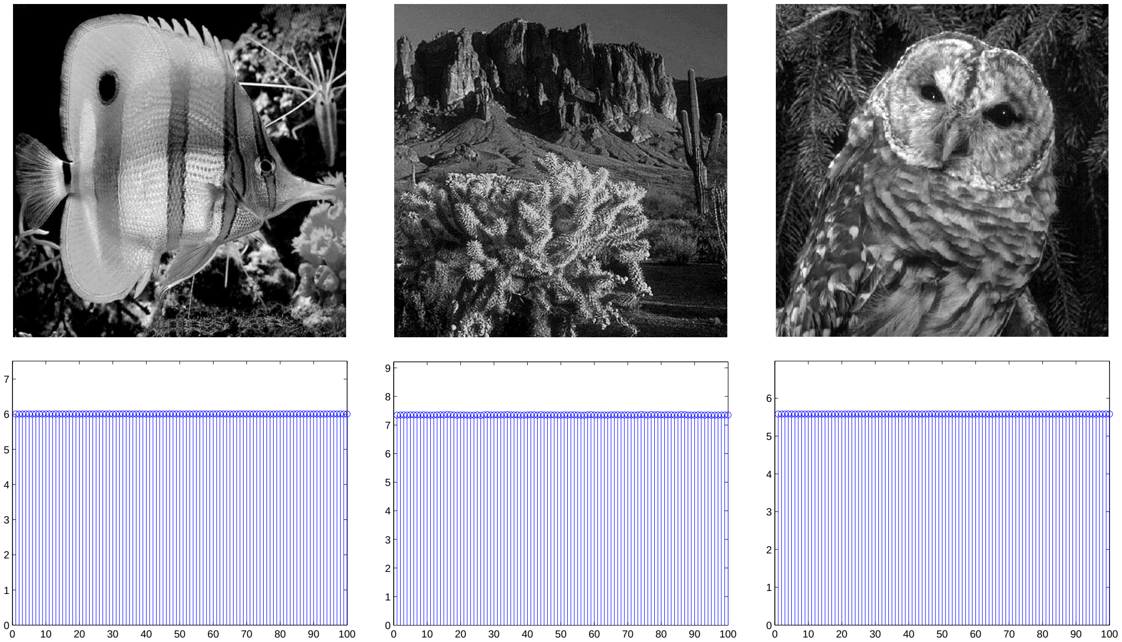


Figure 1.1: Alternating direction method for (1.1.1) **on uncompressed real images seems to always produce the same solution! Top**: Each image is 512×512 in resolution and encoded in the uncompressed `pgm` format (uncompressed images to prevent possible bias towards standard bases used for compression, such as DCT or wavelet bases). Each image is evenly divided into 8×8 non-overlapping image patches (4096 in total), and these patches are all vectorized and then stacked as columns of the data matrix $\boldsymbol{Y}$. **Bottom**: Given each $\boldsymbol{Y}$, we solve (1.1.1) 100 times with independent and randomized (uniform over the orthogonal group) initialization $\boldsymbol{A}_0$. The plots show the values of $\|\boldsymbol{A}_\infty^* \boldsymbol{Y}\|_1$ across the independent repetitions. They are virtually the same and the relative differences are less than 10^{-3}!

instance, regardless of the initialization.

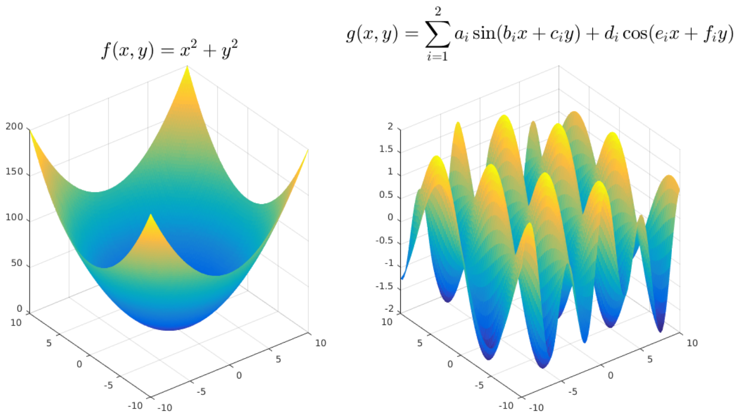
This observation implies the heuristic ADM algorithm may *always converge to a global minimizer*! [5] Equally surprising is that the phenomenon has been observed consistently on real images[6]. One may imagine only generic (e.g., random) data typically have "favorable" structures.

1.2 Nonconvex optimization: theory and practice

The above numerical surprise is a culmination of empirical optimism about nonconvex optimization. Many tasks in applied disciplines can be naturally formulated as nonconvex optimization problems. Simple algorithms, such as gradient descent and alternating direction method, often work surprisingly well in producing good solutions.

[5] Technically, the converge to global solutions is surprising because even convergence of ADM to critical points is atypical, see, e.g., [ABRS10, BST14] and references therein.

[6] Actually the same phenomenon is also observed on simulated data when the coefficient matrix obeys appropriate probability models.


$f(x,y) = x^2 + y^2$

$g(x,y) = \sum_{i=1}^{2} a_i \sin(b_i x + c_i y) + d_i \cos(e_i x + f_i y)$

Figure 1.2: A convex and a nonconvex function in $\mathbb{R}^2$. (Left) $f(x,y) = x^2 + y^2$ is convex; (Right) When a_i through f_i are independent Gaussians, $f(x,y) = \sum_{i=1}^{2} a_i \sin(b_i x + c_i y) + d_i \cos(e_i x + f_i y)$ typically is nonconvex with many spurious local minimizers, besides the global one.

In theory, however, finding global minimizers to nonconvex problems is a daunting task. Generally, even verifying a feasible point is a local minimizer is NP-hard [MK87]. Assuming favorable local geometry and close initialization, local convergence results in optimization typically guarantee that iterative methods produce sequences that converge to a local minimizer [Ber99]. This is obviously not satisfactory as general nonconvex functions can have many spurious local minimizers that are not global (as illustrated in Figure 1.2), and close initialization to even a local minimizer is often unavailable. For general initializations, under technical conditions, global convergence results at best[7] only guarantee that iterative sequences converge to critical points that might be saddle points or local maximizers [Ber99, BAC16], which are undesired.

The gap between theory and practice of nonconvex optimization is evident. The surprising effectiveness of simple algorithms on nonconvex problems lacks a clear explanation. It is tempting to ask what nonconvex problems are easy to solve, and why in practice simple or even heuristic algorithms often succeed in producing high-quality solutions from generic or random initializations.

1.3 Contribution of this thesis

In this thesis, we make a step towards bridging the gap under the hypothesis that

> Certain nonconvex optimization problems have a *benign structure* when the input are *large* and *random/generic*. This benign structure allows "*initialization-free*" iterative methods to *efficiently* find a global minimizer.

[7] ...sequence convergence is not always guaranteed.

Specifically, our contributions include the following:

1. We identify a family of structured nonconvex problems that admit efficient numerical methods for global optimization. This family has a benign geometric structure: (1) All local minimizers are also global; (2) The objective has negative directional curvature around any saddle point or local maximizer. The first property implies absence of spurious local minimizers; the second allows any iterative method that is capable of escaping from saddle points and local maximizers to find a local, and hence global minimizer.

2. We show that this benign geometric structure exists for natural nonconvex formulations of *complete dictionary learning* (CDL) and *generalized phase retrieval* (GPR) under suitable assumptions on the data [SQW15a, SQW16]. These results, together with analogous results established recently on orthogonal tensor decomposition [GHJY15] and noisy phase synchronization and community detection [Bou16, BBV16], underscore the relevance and promise of the geometric structure to central problems in signal processing, machine learning, statistical estimation, and numerous proximal fields. Moreover, the geometric analysis, together with appropriate numerical algorithm, produces novel computational guarantees for these problems. For CDL, we derive the first polynomial-time algorithm for recovering complete dictionaries, when the coefficients have up to linear sparsity (Theorem 6.3). For GPR, we show that generic and reasonably large number of measurements allow recovery of phaseless signals with "initialization-free" numerical methods (Theorem 14.10).

3. The geometric structure facilitates flexible algorithm design. As alluded to above, this benign geometry allows any algorithm with saddle-escaping capability to find a global minimizer, without special initializations. Possibilities include second-order methods such as trust-region method [CGT00, NP06, ABG07, SQW15b], curvilinear search [Gol80], and first-order methods such as noisy/stochastic gradient descent [GHJY15], or even deterministic gradient descent with random initializations [LSJR16]. We describe and analyze second-order trust-region algorithms for CDL and GPR that find the respective global minimizers (up to numerical precision) in polynomial number of iterations, from arbitrary initializations.[8]

Overall, the geometric framework and analysis we develop here for nonconvex optimization proves effective

[8] Qualitatively, other methods we mentioned do not necessarily enjoy the same convergence guarantee. For example, first-order methods cannot ensure convergence to a global minimizer from an arbitrary initialization in a deterministic manner, as saddle point is an attractor for vanilla gradient descent method. Known guarantees either involve randomness in the initialization, or in the iterative step, or both [GHJY15, LSJR16].

on a number of practical problems, and provides a coherent explanation to why "initialization-free" iterative methods can succeed on these problems. The framework and analytic strategy seem to hold promise to many other practical tasks that can be naturally phrased as nonconvex optimization problems.

1.4 Alternative approaches to nonconvex problems

There are numerous generic numerical methods and algorithms developed for tackling nonconvex optimization, mostly centered around the field of global optimization [HPVT00, HP13]. Here we focus on two principled approaches on which recent surge of provable solutions of nonconvex problems is based.

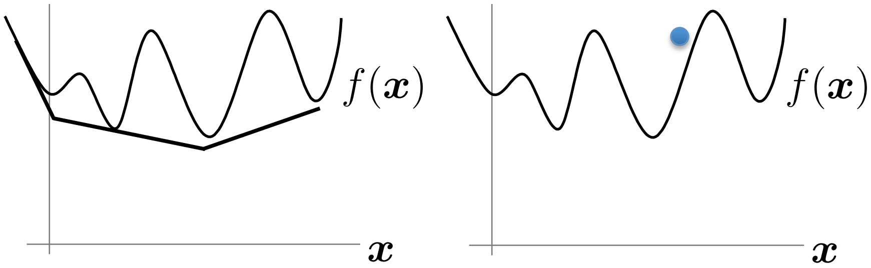


Figure 1.3: Illustration of convex relaxation and initialization plus local refinement approaches to nonconvex problems. (Left) Convex relaxation. A tractable convex surrogate may be hard to find, or suboptimal in performance and expensive in computation. (Right) Initialization plus local refinement. Finding a good initialization poses significant practical challenges.

The first one is convex relaxation, by which one transforms nonconvex problems into convex ones. Roughly speaking, there are two levels at which convex relaxation can occur.

- The basic version (Figure 1.3 (Left)) convexifies the objective function and/or the constraint set in the original space. Natural choices are the convex envelope (or biconjugate)(see, e.g., [HUL93b]) for the objective function and convex hull (see, e.g., [HUL93a]) for the constraint set, which together provide the tightest convex approximation to the original problem but might be computationally intractable[9]. Thus, further or hierarchical relaxations over these are often performed. For the resulting convex problem, one then proves by convex analytic tools that for well structured instances the minimizer is a global minimizer to the original nonconvex problem. This strategy has manifested itself in theoretical and practical advances on recovery of structured signals (e.g., sparse vectors and low-rank matrices, and more [Can14]; see also [CRPW12, Bac10, NYWR09, ALMT14, Tro15a]) in the past decade.

[9] A well-known family of NP-hard convex problems are copositive programming; see, e.g., [Bur12]. Of course, generally convex envelope and convex hull cannot even be conveniently represented algebraically, let alone be computed with.

- The more sophisticated and versatile version involves transformation of the optimization space. For example, Lagrangian dual problems are always concave [Ber99], though they do not always provide useful relaxation to primal problems. Semidefinite programming (SDP) relaxation is a principal approach to deriving approximate solution to quadratic constrained quadratic programs (QCQP) [NWY00], notable results including approximating the MAX-CUT problem [GW95] and solving the trust-region subproblem [RW97]. SDP relaxation also finds numerous applications in combinatorial optimization, control theory (linear matrix inequalities [BEGFB94]), moment problems, among others [NWY00, AL12]. Recently, hierarchical SDP relaxation has been used to approximate sum-of-squares (SOS) optimization, which produces novel computational bounds for a number of nonconvex problems [Par03, Las07, BKS13b, BS14].

Convex relaxation is attractive because it builds on the solid grounds of convex analysis and optimization. However, it does not always ensure correctness, i.e., reproducing the true global minimizer to the originating nonconvex problem, nor even computational tractability. For example, in tensor recovery and nonnegative low-rank approximation, natural convex relaxations are not amenable to efficient computations [HL13, Vav09]. In other cases, although natural convex relaxations are tractable, they are provably suboptimal compared to information-theoretic optimum. Examples include simultaneous structure estimation [OJF$^+$12], tensor recovery [MHWG14], sparse PCA [BR13], and dictionary learning [SWW12].

Another pitfall of convex relaxation is the computational burden it entails. In fact, it is not uncommon that the relaxed problem is a generic SDP, for which generic state-of-art interior-point method based solvers cannot even scale up to medium-size instances. First-order solvers that trade off numerical accuracy for speed (e.g., [OCPB13]) might ameliorate the issue[10], but may still be slow compared to nonconvex alternatives.[11]

The second principled approach involves problem-dependent initializations and subsequent local refinements (Figure 1.3 (Right)). This is the methodology adopted by most recently emerging works on provable nonconvex heuristics. This line of work includes[12] low-rank matrix recovery [KMO10, JNS13, Har14, HW14, NNS$^+$14, JN14, SL14, WCCL15, SRO15, ZL15, TBSR15, CW15], tensor recovery [JO14, AGJ14a, AGJ14b, AJSN15, GHJY15], structured element pursuit [QSW14, HSSS15], dictionary learning [AAJ$^+$13, AGM13, AAN13, ABGM14, AGMM15, SQW15a], mixed regression [YCS13, SA14c], blind deconvolution [LWB13,

[10] In [OCPB13], when the cone of interest is the semidefinite cone, there is a need to perform full eigen-decomposition and projection onto the cone, which is still expensive for large-scale problems.

[11] A notable structured case is when the target solution is known to be low-rank. Then one can apply the factorization trick [BM03] and solve the resulting nonconvex problem. However, the working mechanism is not fully understood; for recent progress, see [BBV16] and the references therein.

[12] See also a webpage maintained by the current author: `http://sunju.org/research/nonconvex/`.

LJ15, LLJB15], generalized phase retrieval [NJS13, CLS15b, CC15, WWS15, SQW16], super resolution [EW15], phase synchronization [Bou16], numerical linear algebra [JJKN15], and so forth. An initialization that is reasonably close to the target solution is evidently critical here; in practice, especially when the attraction basin is small, however, requiring a close initialization almost amounts to solving the problem – which is obviously not easy. Moreover, the "initialization-dependent" theory does not match up empirical observations, in which "initialization-free" algorithms seem to work surprisingly well.

Our approach also follows the nonconvex path, targeting both theoretical soundness and computational practicality. We provide a complete geometric picture of the function landscape; in contrast, assuming close initialization, local refinement analyses are all local in nature. Benign structure of the global geometry allows "initialization-free" iterative methods to solve the nonconvex problems, which sheds light on empirical observations and bear potential to other nonconvex problems we have not covered.

Other provable methods include graduated optimization and reformulation into tensor problems. The former solves a well-constructed sequence of nonconvex problems of increasing complexity to gradually approach the target solution. The technique has proved useful to a number of computer vision and learning problems [BZ87, TC96, MFI15a, MFI15b], but may be very tricky to deploy in general. The latter mostly deals with statistical estimation problems and exploits structure in higher-order moment tensors to extract useful information of the input data [AGH$^+$14, AHJK13, SA14b, SA14a]. Sufficient closeness of finite-sample estimates to their expectations is critical to ensuring stable algorithmic performance, which often incurs large sample complexity.

1.5 Organization

The rest of the thesis is organized as follows. In Chapter 2 we will introduce the $\mathcal{X}$ functions which is the central element of this thesis, provide qualitative and quantitative descriptions, and discuss the trust-region method which can be used as a generic method for minimizing $\mathcal{X}$ functions. We then spend two lengthy parts, Part II and III, presenting two practical problems, complete (sparse) dictionary learning (CDL) and generalized phase retrieval (GPR), that admit natural nonconvex formulations which lie in the $\mathcal{X}$ family. Backgrounds of the problems, detailed geometric characterization of the nonconvex formulations and demonstration they lie in the $\mathcal{X}$ family, and rigorous proof of convergence of the trust-region method tailored to both cases will be presented. We close this thesis (Part IV) by pointing to other important nonconvex problems arising in recent literature that belong to the $\mathcal{X}$ family, and discussing open problems and future directions. Each part or section normally starts with a detailed outline of the local content. The Appendix collects auxiliary results used in proofs in various sections.

Chapter 2

$\mathcal{X}$ Functions

> ...in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity.
>
> R. Tyrrell Rockafellar

Rockafellar is right in separating convex problems from nonconvex ones based on general tractability. However, for most practical problems natural optimization formulations are nonconvex and convex relaxation does not always produce tractable or practical solutions, as we discussed in Section 1.4. On the other hand, there are important nonconvex problems, both classic and new, that we can solve using very simple and natural iterative methods. In this chapter, we start to uncover a benign geometric structure that delineates a family of tractable nonconvex problems. We shall motivate the structure with the classical eigenvector problem (Section 2.1), and then move to a qualitative definition that is convenient for understanding (Section 2.2), and a quantitative definition that is important to provable algorithms (Section 2.3). We will defer novel concrete examples arising from practical problems to the ensuing chapters, but will explain why the geometric structure can be exploited for efficient optimization (Section 2.4) and describe the second-order trust-region method. We will then discuss proof strategy to verifying the geometric structure for particular problems, and to establishing algorithmic convergence based on the trust-region method (Section 2.5). Finally, we close this chapter by touching on implementation issue of the trust-region method towards practicality (Section 2.6). We prefer to make most of our technical statements in Riemannian manifold settings. Concise introduction to the basics can be found in these excellent monographs: [HMG94, Rap97, AMS09].

2.1 The eigenvector problem

The eigenvector problem is fundamental to all applied disciplines. It is considered to be well solved and ready for off-the-shelf applications [GVL12]; however, its natural formulation is nonconvex. For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, consider the Rayleigh quotient formulation for finding the bottom eigenvector:

$$\operatorname{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \ f_{EV}(\boldsymbol{x}) \doteq \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} \quad \text{subject to } \|\boldsymbol{x}\|_2 = 1. \tag{2.1.1}$$

Assume eigenvalues of $\boldsymbol{A}$ are $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{n-1} > \lambda_n$, with the corresponding eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. Problem (2.1.1) is nonconvex because the constraint set $\{\boldsymbol{x} : \|\boldsymbol{x}\|_2 = 1\}$ is the unit sphere in $\mathbb{R}^n$, which is not a convex set. [1]

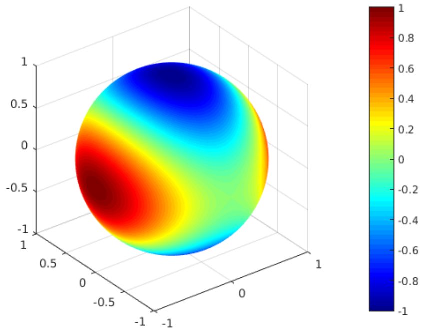


Figure 2.1: Function landscape of a simple eigenvector problem in $\mathbb{R}^3$. In this example, $\boldsymbol{A} = \operatorname{diag}(1, 0, -1)$. Obviously the eigenvalues are $1, 0, -1$, and the corresponding eigenvectors are $\pm\boldsymbol{e}_1$, $\pm\boldsymbol{e}_2$, and $\pm\boldsymbol{e}_3$, which are global maximizers, ridable saddle points, and global minimizers, respectively. Around the saddle points $\pm\boldsymbol{e}_2$, the function has a negative curvature along the $\pm\boldsymbol{e}_3$ direction.

Despite the nonconvexity, one can explicitly locate the critical points of (2.1.1): they are exactly the signed eigenvectors, $\pm\boldsymbol{v}_1, \ldots, \pm\boldsymbol{v}_n$. Moreover, by examining the second-order geometry[2], we can classify critical points of f_{EV} more precisely (see. e.g., Proposition 4.6.2 of [AMS09]):

> The only global minimizers are $\pm\boldsymbol{v}_n$; the only global maximizers are $\pm\boldsymbol{v}_1$; the intermediate eigenvectors $\boldsymbol{v}_k$'s (i.e., the remaining critical points) are saddle points, around each there is a negative curvature in $\pm\boldsymbol{v}_n$ direction.

Figure 2.1 shows the landscape of f_{EV} for $\boldsymbol{A} = \operatorname{diag}(1, 0, -1)$, for which obviously $\pm\boldsymbol{e}_3$ are the global minimizers, and $\pm\boldsymbol{e}_2$ are the saddle points. Note that around $\pm\boldsymbol{e}_2$, f_{EV} has a negative curvature in the $\pm\boldsymbol{e}_3$ direction.

[1] In fact, when treated as a function defined on the sphere, a particular Riemannian manifold, the function is not convex in geodesic sense either. The reason is that the two only global minimizers $\boldsymbol{v}_n$ and $-\boldsymbol{v}_n$ are isolated on the sphere, while geodesic convexity implies the set of global minimizers be geodesically totally convex – similar to the Euclidean-space results; see, e.g., Section 3.7 of [Udr94] or Section 6.2 of [Rap97].

[2] One can either introduce a Lagrange multiplier and verify the second-order sufficient conditions (see, e.g., Sections 3.1 & 3.2 of [Ber99]), or treat the function as defined on the sphere and resort to Riemannian gradient and Hessian (see, e.g., Section 4.6 of [AMS09]) for the analysis.

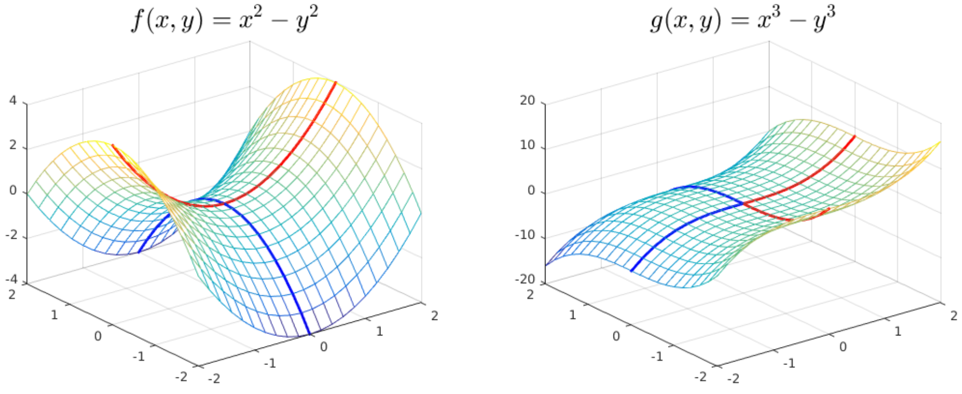
The landscape of f_{EV} on the sphere differs significantly from our typical "mental picture" of nonconvex functions as possessing many spurious local minimizers – here all local minimizers are also global. The negative directional curvature around each saddle point of f_{EV} is also important for analysis and computation – we shall explain this in the next sections. The successes of efficient iterative methods to compute eigenvalues and eigenvectors of general symmetric matrices hinge on these nice geometric structures; this deep connection becomes evident when these iterative methods are interpreted from the Riemannian optimization standpoint, see, e.g., relevant chapters of [HMG94, AMS09].

2.2 $\mathcal{X}$ functions: qualitative version

Inspired by the eigenvector problem (chronologically, by the sparse dictionary learning problem of Part II first), we single out a class of nonconvex problems that seems structured enough to admit efficient optimization methods.

> **Definition 2.1 ($\mathcal{X}$ functions: qualitative version I)** *Let $\mathcal{M}$ be a Riemannian manifold. A twice continuously differentiable function $f : \mathcal{M} \mapsto \mathbb{R}$ is said to be in the $\mathcal{X}$ class if:*
>
> *(P-1) All local minimizers of f are also global minimizers;*
>
> *(P-2) All saddle points of f have a negative directional curvature.*

(P-1) rules out the presence of spurious local minimizers other than the global ones; (P-2) may appear extraneous to many and deserves more elaboration. For simplicity, consider a twice continuously differentiable


$f(x, y) = x^2 - y^2$

$g(x, y) = x^3 - y^3$

Figure 2.2: Illustration of ridable saddles vs. generic saddles. At a ridable saddle, the Hessian has a negative eigenvalue, whereas general saddles may be shaped by higher-order derivatives in directions where second-order derivatives vanish. Shown in the plot are functions $f(x, y) = x^2 - y^2$ (Left) and $g(x, y) = x^3 - y^3$ (Right). For g, both first- and second-order derivatives vanish at $(0, 0)$, and the local function landscape is determined by the third-order derivatives. In both plots, red curves indicate local ascent directions and blue curves indicate local descent directions.

function $f : \mathbb{R}^n \mapsto \mathbb{R}$. Critical points are points $\boldsymbol{x} \in \mathbb{R}^n$ such that $\nabla f(\boldsymbol{x}) = \mathbf{0}$. By definition, saddle points of f are those critical points which are neither local maximizers nor local minimizers. In other words, a critical point $\boldsymbol{x}$ is a saddle point of f if for every $\varepsilon > 0$, the ε-neighborhood of $\boldsymbol{x}$ contains simultaneously an $\boldsymbol{x}_-$ such that $f(\boldsymbol{x}_-) < f(\boldsymbol{x})$ and an $\boldsymbol{x}_+$ such that $f(\boldsymbol{x}_+) > f(\boldsymbol{x})$. While saddle points are typically illustrated in textbooks as being literally saddle shaped (Figure 2.2, Left), they do not necessarily be so – particularly, around saddle points the function f does not necessarily have a negative curvature in any direction (Figure 2.2, Right). Technically, textbook saddle points are those whose Hessian has a negative eigenvalue, which induces a negative curvature direction; we shall call these prototypical saddles as *ridable saddles*.

> **Definition 2.2 (Ridable saddles; also strict saddles in [GHJY15])** *Let $\mathcal{M}$ be a Riemannian manifold and consider a twice continuously differentiable function $f : \mathcal{M} \mapsto \mathbb{R}$. A point $\boldsymbol{x} \in \mathcal{M}$ is said to be a ridable saddle point if its Riemannian Hessian* $\operatorname{Hess} f(\boldsymbol{x})$ *has a negative eigenvalue on $T_{\boldsymbol{x}}\mathcal{M}$, i.e., the tangent space of $\mathcal{M}$ at $\boldsymbol{x}$.*

By comparison, other types of saddles arise only when the Hessian is positive semidefinite (possibly $\mathbf{0}$) at a critical point such that there exists a direction in which the second-order variation also vanishes. In this case, the directional function landscape is determined by higher-order information, which renders the current critical point either a local minimizer, a local maximizer, or an unridable saddle point.

The reason we favor ridable saddles is that there is a natural descent direction that can be exploited efficiently by iterative methods; we shall detail on this in the ensuing sections. For the sake of performing numerical optimization, we will augment (P-2) above slightly, extending it into a characterization of all critical points.

> **Definition 2.3 ($\mathcal{X}$ functions: qualitative version II)** *Let $\mathcal{M}$ be a Riemannian manifold. A twice continuously differentiable function $f : \mathcal{M} \mapsto \mathbb{R}$ is said to be in the $\mathcal{X}$ class if:*
>
> *(P-1) All local minimizers of f are also global minimizers;*
>
> *(P-2A) For all local minimizers,* $\operatorname{Hess} f \succ \mathbf{0}$*;*
>
> *(P-2B) For all other critical points, $\lambda_{\min}(\operatorname{Hess} f) < 0$.*
>
> *Moreover, f is said to be a ridable-saddle function (also strict-saddle function [GHJY15]) if (P-2A) and (P-2B) hold.*

(P-2A), together with smoothness, implies the function is locally strongly convex around any local minimizer. This is a typical assumption to be made even to establish a local convergence result. (P-2B) requires all saddle points be ridable saddles, and also around each local maximizer f has a negative directional curvature. The latter precludes the existence of a full-dimensional plateau of local maximizers – where an interior local maximizer does not admit any local descent direction.

One might think (P-2A) and (P-2B) are artificial and possibly demanding, but in fact ridable functions are prevalent. A smooth function $f : \mathcal{M} \mapsto \mathbb{R}$ is called *Morse* if all critical points are non-degenerate, i.e., with full-rank Hessians. It is easy to see that *all Morse functions are ridable-saddle functions*. The abundance of Morse functions is a fundamental result in Morse theory, which says the Morse functions form an open, dense subset of all smooth functions $\mathcal{M} \mapsto \mathbb{R}$– a generic smooth function is Morse![3]

2.3 $\mathcal{X}$ functions: quantitative version

The qualitative definition for $\mathcal{X}$ functions (Definition 2.3) carves out the structured nonconvex problems of interest, but is not adequate for numerical computation and algorithmic analysis. It lacks in two aspects: (1) description of noncritical points. Numerical optimization methods mostly entail iterative sequences that move across numerous noncritical points before finally converging to a critical point; (2) quantitative description of critical points. For example, (P-2A) does not prevent $\operatorname{Hess} f$ from being arbitrarily close to positive semidefinite, which leads to dramatically different algorithmic behaviors for iterative methods as compared to strongly definite cases. (P-2B) suffers from similar deficiency.

We now provide a quantitative definition for $\mathcal{X}$ functions that alleviates the above problems and provides concrete workable quantities.

> **Definition 2.4 ($\mathcal{X}$ functions: quantitative version)** *Let $\mathcal{M}$ be a Riemannian manifold. A twice continuously differentiable function $f : \mathcal{M} \mapsto \mathbb{R}$ is said to be in the $\mathcal{X}$ class with parameter $(\alpha, \beta, \delta, \gamma)$ $(\alpha, \beta, \delta, \gamma > 0)$ if:*
>
> *(**No spurious minimizers**) All local minimizers of f are also global minimizers,*
>
> *and f is a $(\alpha, \beta, \gamma, \delta)$ ridable-saddle function (also strict-saddle function [GHJY15] [4]) defined as follows: any point $\boldsymbol{x} \in \mathcal{M}$ obeys at least one of the following:*
>
> *(**Strong gradient**) $\|f(\boldsymbol{x})\| \geq \beta$;*
>
> *(**Negative curvature**) There exists $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{M}$ with $\|\boldsymbol{v}\| = 1$ such that $\langle \operatorname{Hess} f(\boldsymbol{x})[\boldsymbol{v}], \boldsymbol{v} \rangle \leq -\alpha$;*
>
> *(**Strong convexity around minimizers**) There exists a local minimizer $\boldsymbol{x}_\star$ such that $\|\boldsymbol{x} - \boldsymbol{x}_\star\| \leq \delta$, and for all $\boldsymbol{y} \in \mathcal{M}$ with $\|\boldsymbol{x}_\star - \boldsymbol{y}\| \leq 2\delta$, $\langle \operatorname{Hess} f(\boldsymbol{y})[\boldsymbol{v}], \boldsymbol{v} \rangle \geq \gamma$ for any $\boldsymbol{v} \in T_{\boldsymbol{y}}\mathcal{M}$ with $\|\boldsymbol{v}\| = 1$.*

[3] See, e.g., Section 2.2.c of [Mat02], or Section 1.2 of [Nic11] for the precise statement and proof. In fact, Section 2.2.c of [Mat02] proves the result by showing an arbitrarily small generic linear perturbation to a non-Morse function turns the function into a Morse function.

[4] When $\mathcal{M}$ is $\mathbb{R}^n$ or $\mathbb{C}^n$, the two definitions coincide. It is interesting to see if the two agree in general settings. Particularly, [GHJY15]

The “no spurious minimizers” condition is simple and inherits from the qualitative version. To define a quantitative ridable-saddle function, we divide the Riemannian manifold $\mathcal{M}$ into two subsets: one is far apart from any critical point, signified by having gradients with large magnitudes (“strong gradient”, characterized by β), and one consists of neighborhoods of critical points. We further divide the second into two subsets: one consists of neighborhoods of local maximizers or saddle points where at each point the function has a negative directional curvature (“negative curvature”, characterized by α), and one consists of neighborhoods of local minimizers where the function is locally strongly convex around each local minimizer (“strongly convexity around minimizers", characterized by γ, δ).

2.4 Algorithm: second-order trust-region method

$\mathcal{X}$ functions have no spurious local minimizers, and thus finding any local minimizer solves the associated minimization problems. Without assuming special initializations, typical iterative methods only guarantee to converge to critical points at best, which might be saddles points or local maximizers that obviously are not desired. So the central task for iterative methods here is avoiding being trapped by saddle points and local maximizers.

A number of iterative methods are empirically observed to be immune to saddle points and local maximizers, with varied theoretical guarantees and empirical efficiencies. These methods include second-order methods such as trust-region method [CGT00, NP06, ABG07, SQW15b], curvilinear search [Gol80], and first-order methods such as noisy/stochastic gradient descent [GHJY15], or even deterministic gradient descent with random initializations [LSJR16]. In this thesis we focus on the second-order trust-region method that produces strong computational guarantees for minimizing $\mathcal{X}$ functions (polynomial-time algorithm without special initializations[5]) with relatively simple proof.[6]

The reason second-order methods can help escape from ridable saddle points or local maximizers with negative direction curvature is simple. Consider, for simplicity, a twice continuously differentiable function

deals only with manifold defined by equalities of the form $c_i(\boldsymbol{x}) = 0$ with differentiable function c, which excludes many manifolds of interest, such as symmetric positive definite matrices of a fixed dimension. See this page: `http://www.manopt.org/tutorial.html#manifolds` for more examples. See also discussion in Introduction of this paper [ABG07] on the (incompatible) relationship between manifold optimization and constrained optimization in the Euclidean space.

[5] ... provided the $\mathcal{X}$ parameters and smoothness parameters of the problem at hand are treated as constant, or if not, their dependency on the problem size is “reasonable” – α, β, γ are reasonably large (say bounded from below by inverse polynomial of problem size) and Lipschitz constants are reasonably small (say bounded above by polynomials of problem size).

[6] We believe that the curvilinear search method [Gol80] very likely has similar computational guarantees on $\mathcal{X}$ functions. Since we are mostly concerned with understanding the $\mathcal{X}$ property of practical nonconvex problems, we will not pursue a rigorous analysis of this method in this thesis.

$f : \mathbb{R}^n \mapsto \mathbb{R}$ and a saddle point $\boldsymbol{x}_0 \in \mathbb{R}^n$ of f. Around $\boldsymbol{x}_0$, the second-order Taylor approximation to f is:

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}_0) = f(\boldsymbol{x}_0) + \frac{1}{2}\boldsymbol{\delta}^\top \nabla^2 f(\boldsymbol{x}_0)\boldsymbol{\delta}.$$

If $\nabla^2 f(\boldsymbol{x}_0)$ has a negative eigenvalue λ_- with its corresponding eigenvector $\boldsymbol{v}_-$, setting $\delta_0 = t\boldsymbol{v}_-$ leads to

$$\widehat{f}(\boldsymbol{\delta}_0; \boldsymbol{x}_0) - f(\boldsymbol{x}_0) = -\frac{1}{2}t^2 \left|\lambda_-\right| < 0.$$

When t is reasonably small such that the second-order Taylor approximation of $\widehat{f}$ to f is reasonably accurate (i.e., higher-order information is negligible) within a t-neighborhood of $\boldsymbol{x}_0$, moving in $\boldsymbol{v}_-$ direction induces a strict decrease to the function value. In other words, $\boldsymbol{v}_-$ is a local descent direction. Analogous argument explains why local maximizers with a negative directional curvature are not attractors of second-order methods.

For nonconvex problems, the Hessian $\nabla^2 f(\boldsymbol{x})$ is not always definite. A natural way of dealing with both definite and indefinite Hessians consistently is performing optimization within locally restricted regions – calling to the *trust-region method* (TRM). For a function $f : \mathbb{R}^n \to \mathbb{R}$ and an unconstrained optimization problem

$$\mathrm{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \; f\left(\boldsymbol{x}\right),$$

typical (second-order) TRM proceeds by successively forming quadratic approximations to f at the current iterates,

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}^{(k-1)}) \doteq f(\boldsymbol{x}^{(k-1)}) + \nabla^\top f(\boldsymbol{x}^{(k-1)})\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{Q}(\boldsymbol{x}^{(k-1)})\boldsymbol{\delta}, \tag{2.4.1}$$

where $\boldsymbol{Q}\left(\boldsymbol{x}^{(k-1)}\right)$ is a proxy for the Hessian matrix $\nabla^2 f\left(\boldsymbol{x}^{(k-1)}\right)$, which encodes the second-order geometry. The next movement direction is determined by seeking a minimizer of $\widehat{f}\left(\boldsymbol{\delta}; \boldsymbol{x}^{(k-1)}\right)$ over a small region, normally a norm ball $\left\|\boldsymbol{\delta}\right\|_p \le \Delta$, called the *trust region*, inducing the well studied trust-region subproblem:

$$\boldsymbol{\delta}^{(k)} \doteq \operatorname*{arg\,min}_{\boldsymbol{\delta} \in \mathbb{R}^n, \left\|\boldsymbol{\delta}\right\|_p \le \Delta} \widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}^{(k-1)}), \tag{2.4.2}$$

where Δ is called the trust-region radius that controls how far the movement can be made. Norms $p = 1, 2, \infty$ and their rescaled versions are often used in practice; we focus on $p = 2$ since efficient algorithms exist to solve (2.4.2) in this case[7], see e.g., [MS83, RW97, CGT00, FW04, HK14]. Once $\boldsymbol{\delta}^{(k)}$ is found, the next iterate is

[7] For the $p = 1, \infty$ cases, there are no known efficient algorithms to solve them globally, though the problem is considerably ameliorated in practice as reasonable approximate solutions to (2.4.2) still suffice to guarantee convergence of TRM; see relevant discussions in Section 7.8 of [CGT00].

determined by the updating formula

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + \boldsymbol{\delta}^{(k)}.$$

Detailed introductions to the classical TRM can be found in the texts [CGT00, NW06].

TRM can naturally be adapted for optimization over Riemannian manifolds [ABG07, AMS09]:

$$\operatorname{minimize}_{\boldsymbol{x}\in\mathcal{M}} f(\boldsymbol{x}). \tag{2.4.3}$$

The canonical quadratic approximation at the current iterate now takes the form

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}^{(k-1)}) \doteq f(\boldsymbol{x}^{(k-1)}) + \left\langle \operatorname{grad} f(\boldsymbol{x}^{(k-1)}), \boldsymbol{\delta} \right\rangle + \frac{1}{2} \left\langle \operatorname{Hess} f(\boldsymbol{x}^{(k-1)})[\boldsymbol{\delta}], \boldsymbol{\delta} \right\rangle, \tag{2.4.4}$$

where $\operatorname{grad} f(\boldsymbol{x})$ and $\operatorname{Hess} f(\boldsymbol{x})$ are the Riemannian gradient and Riemannian Hessian of f at point $\boldsymbol{x}$, acting on $T_{\boldsymbol{x}}\mathcal{M}$ and the approximation is defined for any $\boldsymbol{\delta} \in T_{\boldsymbol{x}}\mathcal{M}$; related concepts are briefly illustrated in Figure 2.3. The trust-region subproblem becomes

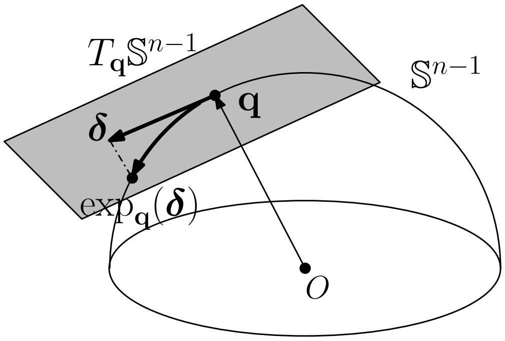

$T_{\mathbf{q}}\mathbb{S}^{n-1}$
$\mathbb{S}^{n-1}$
$\mathbf{q}$
$\boldsymbol{\delta}$
$\exp_{\mathbf{q}}(\boldsymbol{\delta})$
O

Figure 2.3: Illustrations of the tangent space and exponential map, and explanation of quadratic approximation for functions over Riemannian manifolds. Here the exemplar manifold is the sphere $\mathbb{S}^{n-1}$. Let $\mathbb{B}(\boldsymbol{q}, \varepsilon)$ be the ball centered at $\boldsymbol{q}$ with radius ε. Exponential map is a canonical way of "pulling" a vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathcal{M} \cap \mathbb{B}(\boldsymbol{q}, \varepsilon)$ to $\mathcal{M}$ for small ε.[8]One can define a function $\overline{f}(\boldsymbol{\delta}; \boldsymbol{q})$: $T_{\boldsymbol{q}}\mathcal{M} \cap \mathbb{B}(\boldsymbol{q}, \varepsilon) \mapsto \mathbb{R}$ as $\overline{f}(\boldsymbol{\delta}; \boldsymbol{q}) \doteq f \circ \exp_{\boldsymbol{q}}(\boldsymbol{\delta})$. Then quadratic approximation (2.4.4) is just the second-order Taylor expansion of $\overline{f}(\boldsymbol{\delta}; \boldsymbol{q})$ around the point $\boldsymbol{q} \doteq \boldsymbol{x}^{(k-1)}$.

$$\boldsymbol{\delta}^{(k)} \doteq \mathop{\arg\min}_{\boldsymbol{\delta}\in T_{\boldsymbol{x}^{(k-1)}}\mathcal{M}, \|\boldsymbol{\delta}\|_2 \le \Delta} \widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}^{(k-1)}). \tag{2.4.5}$$

Most manifolds of practical interest are embedded submanifolds of $\mathbb{R}^{m\times n}$ and the tangent space is a subspace of $\mathbb{R}^{m\times n}$. For an $\boldsymbol{x}^{(k-1)} \in \mathcal{M}$ and an orthonormal basis $\boldsymbol{U}$ for $T_{\boldsymbol{x}^{(k-1)}}\mathcal{M}$, one can solve (2.4.5) by solving the recast Euclidean trust-region subproblem

$$\boldsymbol{\xi}^{(k)} \doteq \mathop{\arg\min}_{\|\boldsymbol{\xi}\|_2 \le \Delta} \widehat{f}(\boldsymbol{U}\boldsymbol{\xi}; \boldsymbol{x}^{(k-1)}). \tag{2.4.6}$$

Then we have $\boldsymbol{\delta}^{(k)} = \boldsymbol{U}\boldsymbol{\xi}^{(k)}$. The point $\boldsymbol{x}^{(k-1)} + \boldsymbol{\delta}^{(k)}$ lies in $T_{\boldsymbol{x}^{k-1}}\mathcal{M}$, but generally it is not on $\mathcal{M}$. We need

[8] Generally exponential map is only locally defined in a neighborhood around the origin of the tangent space. It suffices for our purposes however, as we only care about local approximation of the function.

to pull it back to the manifold by performing a *retraction* step $R_{\boldsymbol{x}^{(k-1)}}(\cdot)$ (e.g., by exponential map for any Riemannian manifold, or Euclidean projection for embedded submanifolds of Euclidean spaces [AM12]). Hence the update formula reads

$$\boldsymbol{x}^{(k)} = R_{\boldsymbol{x}^{(k-1)}}(\boldsymbol{x}^{(k-1)} + \boldsymbol{\delta}^{(k)}). \tag{2.4.7}$$

It is possible to run TRM on general constrained problems beyond the Riemannian manifold setting; see, e.g., [CGT00]. To derive provable algorithms for these problems, second-order geometry of Lagrangian function plays a pivotal role. [GHJY15] deals with problems with only equality constraints.

2.5 Sketch of proof ideas

We have described the (Riemannian) second-order trust-region method that is powerful enough to escape from saddle points and local maximizers of $\mathcal{X}$ functions and finally find a local/global minimizer of $\mathcal{X}$ functions, from arbitrary initializations. Thus, the key challenge is how to determine whether a given function lies in the $\mathcal{X}$ family. We will provide an overview of the strategy we have developed in our works [SQW15a, SQW16, SQW15b] – details occupy main bodies of next two parts. Then, we will discuss how to establish concrete convergence rates of the trust-region algorithms. Applying these ideas to our problems [SQW15a, SQW16] involves local adaptation of definition of $\mathcal{X}$ functions and the trust-region method.

2.5.1 Identifying $\mathcal{X}$ functions

To recognize an $\mathcal{X}$ function, a natural idea is to analytically locate the critical points and then analyze their Hessians. This may work for very simple cases (e.g., low-order polynomials[9]), but will fail badly in general as locating critical points itself may already involve algebraic equations that are not amenable to analytic techniques[10]. Drawing analog from convex analysis, it appears the best one can hope is to figure out operational rules that preserve the $\mathcal{X}$-ness such that complex $\mathcal{X}$ functions can be constructed from simple ones – this likely requires intensive further research efforts.

[9] One classical example is the eigenvector problem we presented in Section 2.1; see, e.g., Section 4.6 of [AMS09] for proof; see also discussion in [SQW15b].

[10] In fact, if explicit locating critical points could be done analytically in general, one can solve all bounded optimization problems with polynomial numbers of critical points efficiently.

In this thesis, we focus on nonconvex problems of the form

$$\operatorname{minimize}_{\boldsymbol{x}} F(\boldsymbol{x}) \doteq \frac{1}{m}\sum_{k=1}^{m} f(\boldsymbol{x}; \boldsymbol{y}_k) \quad \text{subject to} \quad \boldsymbol{x} \in \mathcal{M}, \tag{2.5.1}$$

where $\mathcal{M}$ is a Riemannian manifold and $\{\boldsymbol{y}_k\}$ is an ensemble of observations that often bears randomness. This framework is powerful enough to encompass many problems arising in signal processing and machine learning, typically in recovery of structured signals where f is a penalty to promote the desired structure in the solution; here we will briefly describe *complete (sparse) dictionary learning* (CDL) and *generalized phase retrieval* (GPR) as concrete examples – again, substantial amounts of details are included in the next two parts; more practical examples appearing in the literature will be discussed in Chapter 19.

> **Example 2.5 (Complete (Sparse) Dictionary Learning [SQW15a] (CDL))** *Arising in signal processing and machine learning, dictionary learning tries to approximate a given data matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ as the product of a dictionary $\boldsymbol{A}$ and a sparsest coefficient matrix $\boldsymbol{X}$. In recovery setting, assuming $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$ with $\boldsymbol{A}_0$ complete (square and invertible), $\boldsymbol{Y}$ and $\boldsymbol{X}_0$ have the same row space. Under appropriate (probabilistic) model on $\boldsymbol{X}_0$, it makes sense to recover one row of $\boldsymbol{X}_0$ each time by finding the sparsest direction*[11] *in* $\operatorname{row}(\boldsymbol{Y})$ *through the optimization:*
>
> $$\operatorname{minimize}_{\boldsymbol{q}} \left\|\boldsymbol{q}^\top \boldsymbol{Y}\right\|_0 \quad \text{subject to} \quad \boldsymbol{q} \neq \boldsymbol{0},$$
>
> *which can be relaxed as a nonconvex problem*
>
> $$\operatorname{minimize} F(\boldsymbol{q}) \doteq \frac{1}{p}\sum_{k=1}^{p} h(\boldsymbol{q}^\top \overline{\boldsymbol{y}}_k) \quad \text{subject to} \quad \|\boldsymbol{q}\|_2 = 1 \quad [\textit{i.e.}, \boldsymbol{q} \in \mathbb{S}^{n-1}]. \tag{2.5.2}$$
>
> *Here $h(\cdot)$ is a smooth approximation to the $|\cdot|$ function which promotes sparsity and $\overline{\boldsymbol{y}}_k$ is the k-th column of $\overline{\boldsymbol{Y}}$, a proxy of $\boldsymbol{Y}$. The manifold $\mathcal{M}$ is $\mathbb{S}^{n-1}$ here.*

> **Example 2.6 (Generalized Phase Retrieval [SQW16] (GPR))** *For complex signal $\boldsymbol{x} \in \mathbb{C}^n$, generalized phase retrieval (PR) tries to recover $\boldsymbol{x}$ from nonlinear measurements of the form $y_k = |\boldsymbol{a}_k^* \boldsymbol{x}|$, for $k = 1, \dots, m$, where $\{\boldsymbol{a}_k\}$ is a set of known random complex vectors (say i.i.d. sub-Gaussian). This task has occupied the central place in imaging systems for scientific discovery [SEC+15], among many others. Assuming i.i.d. Gaussian*

[11] The absolute scale is not recoverable.

> *measurement noise, a natural formulation for GPR is*
>
> $$\operatorname{minimize}_{\boldsymbol{z}\in\mathbb{C}^n} F(\boldsymbol{z}) \doteq \frac{1}{4m}\sum_{k=1}^{m}\left(y_k^2 - |\boldsymbol{a}_k^*\boldsymbol{z}|^2\right)^2, \tag{2.5.3}$$
>
> *which is a* 4*-th order polynomial and is nonconvex. The manifold* $\mathcal{M}$ *here is* $\mathbb{C}^n$*.*

One salient feature about the above recovery problems is that there are clear objects of interest to be recovered as global minimizers of the respective nonconvex formulations. According to Definition 2.4, to show the nonconvex formulations lie in the $\mathcal{X}$ family, it is natural to show the nonconvex functions are locally strongly convex around the target minimizers but have no other minimizers (by demonstrating either strong gradients or negative directional curvatures) outside these neighborhoods. Specifically, one can start with verifying the ridable-saddle property:

> ***Verifying the ridable-saddle property****: Partition* $\mathcal{M}$ *into three regions, corresponding to the strong gradient, negative curvature, and local strongly convex regions in Definition 2.4 respectively, and check the respective associated quantities.*

How to divide the regions outside the vicinity of the target minimizers into gradient and curvature regions is tricky and problem-dependent. One coupled challenge is acquiring knowledge of directions with negative curvature when the gradient is weak. Our current strategy is first gaining intuition from low-dimensional plot of the function landscape and then trying to extrapolate the visual structural division to high dimensions. To complete the geometric characterization, one then show that all local minimizer are also global:

> ***Verifying all local minimizers are global****: Show all target minimizers are surrounded by a local strongly convex region and verify the targets are indeed local minimizers, say, by checking gradients and all target minimizers carry the same function value.*

That each target minimizer has a strongly convex neighborhood should already be accounted for in the above explicit division of $\mathcal{M}$ into three regions. The target minimizers are the only minimizers, as each local strongly convex region can have at most one local minimizer.

Derivatives of (2.5.1) likely are also sum of random quantities when the input data $\{\boldsymbol{y}_k\}$ are random. Thus, it is natural to follow a typical expectation-concentration path for the analysis [BLM13]: first demonstrate the expected version of the function verifies the $\mathcal{X}$ property by working with the expected derivatives, and then show the various derivative terms in the divided regions concentrate well around their respective expectations, when m is reasonably large. Moreover, to show the gradient is strong in the strong gradient region, it suffices to show a directional derivative is strong – the latter again entails choosing a direction to work on, but showing scalar concentration is often easier than the vector counterpart.

2.5.2 Proving convergence of trust-region method

For simplicity, we assume the trust-region size parameter Δ is fixed to a small value. This ensures that local quadratic approximations model the local behaviors of the function reasonably accurately, such that the qualitative effect of a local movement on the objective can be gauged from that on the local approximation. We also assume trust-region subproblems are solved exactly.

Based on the above idealizations, each step at a negative-curvature or strong-gradient point decreases the objective value by a concrete amount. To see the reason, consider the Euclidean case for simplicity. When gradient is strong, taking $\boldsymbol{\delta} = -\Delta \cdot \nabla f(\boldsymbol{x})$ leads to

$$\begin{aligned} \widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}) &= f(\boldsymbol{x}) - \Delta \left\| \nabla f(\boldsymbol{x}) \right\|_2^2 + \frac{1}{2} \Delta^2 \left(\nabla f \right)^\top \nabla^2 f(\boldsymbol{x}) \nabla f \\ &\le f(\boldsymbol{x}) - \Delta \left\| \nabla f(\boldsymbol{x}) \right\|_2^2 + \frac{1}{2} \Delta^2 \left\| \nabla f(\boldsymbol{x}) \right\|_2^2 \left\| \nabla^2 f(\boldsymbol{x}) \right\| . \end{aligned}$$

Hence Δ is sufficiently small, $\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}) - f(\boldsymbol{x}) < 0$. On the other hand, when the curvature is negative in a certain direction, say $\boldsymbol{\delta}_0$, with curvature parameter α_0, then taking

$$\boldsymbol{\delta} = \begin{cases} -\Delta \frac{\operatorname{sign}\left(\boldsymbol{\delta}_0^\top \nabla f(\boldsymbol{x}) \right) \boldsymbol{\delta}_0}{\left\| \operatorname{sign}\left(\boldsymbol{\delta}_0^\top \nabla f(\boldsymbol{x}) \right) \boldsymbol{\delta}_0 \right\|_2} & \nabla f(\boldsymbol{x}) \neq \mathbf{0}, \\ \Delta \frac{\boldsymbol{\delta}_0}{\left\| \boldsymbol{\delta}_0 \right\|_2} & \nabla f(\boldsymbol{x}) = \mathbf{0} \end{cases}$$

leads to

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}) \le f(\boldsymbol{x}) - \frac{1}{2} \alpha_0 \Delta^2.$$

Thus, with additional assumptions such as compact sub-level set on the function, we conclude that the iterate sequence finally moves into the strongly convex neighborhood of a minimizer. Then a trust-region step is either constrained (i.e., the trust-region constraint is active) such that it also deceases the objective by a concrete amount (the reason is the same as that for the strong gradient point), or unconstrained, which is a good indicator that the target minimizer is within a radius Δ. In the latter case, the algorithm behaves much like the classical Newton method and quadratic sequence convergence can be shown. [SQW15a, SQW16] include detailed arguments for our two examples.

2.6 Running trust-region algorithms in practice

Fixing a small step size and solving trust-region subproblem exactly ease the analysis, but also render the TRM algorithm impractical. In practice, the trust-region subproblem is never exactly solved, and the trust-region step size is adjusted to the local geometry, say by backtracking. It is possible to modify our algorithmic analysis to account for inexact subproblem solvers and adaptive step size; for sake of brevity, we do not pursue it in this thesis. Recently, [BAC16] has made progress towards this direction.

We close this chapter by introducing the *Manopt* toolbox [BMAS14][12]. Manopt is a user-friendly Matlab toolbox that implements several sophisticated solvers for tackling optimization problems over Riemannian manifolds. The most developed solver is based on the TRM. This solver uses the truncated conjugate gradient (tCG; see, e.g., Section 7.5.4 of [CGT00]) method to (approximately) solve the trust-region subproblem (vs. the exact solver in our analysis). It also dynamically adjusts the step size using backtracking. However, the original implementation (Manopt 2.0) is not adequate for our purposes. Their tCG solver uses the gradient as the initial search direction, which does not ensure that the TRM solver can escape from saddle points [ABG07, AMS09]. We modify the tCG solver, such that when the current gradient is small and there is a negative curvature direction (i.e., the current point is near a saddle point or a local maximizer for $f(z)$), the tCG solver explicitly uses the negative curvature direction[13] as the initial search direction. This modification ensures the TRM solver always escapes saddle points/local maximizers with negative directional curvature. Hence, the modified TRM algorithm based on Manopt is expected to have the same qualitative behavior as the idealized version we described above, with better scalability. We will perform our numerical simulations using the modified TRM algorithm whenever necessary.

[12] Available online: `http://www.manopt.org`.

[13] ...adjusted in sign to ensure positive correlation with the gradient – if it does not vanish.

Part II

Complete (Sparse) Dictionary Learning

In this part, we consider the problem of recovering a complete (i.e., square and invertible) matrix $\boldsymbol{A}_0$, from $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$ with $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0$, provided $\boldsymbol{X}_0$ is sufficiently sparse. This recovery problem is central to the theoretical understanding of dictionary learning, which seeks a sparse representation for a collection of input signals and finds numerous applications in modern signal processing and machine learning. We give the first efficient algorithm that provably recovers $\boldsymbol{A}_0$ when $\boldsymbol{X}_0$ has $O(n)$ nonzeros per column, under suitable probability model for $\boldsymbol{X}_0$. In contrast, prior efficient algorithms either only guarantee recovery in the super-sparse regime ($O(\sqrt{n})$ nonzeros per column in $\boldsymbol{X}_0$), or require solving multiple rounds of sum of squares to attain the $O\left(n^{1-\delta}\right)$ regime for any constant $\delta \in (0,1)$, which is only of theoretical interest.

Our algorithm centers around solving a nonconvex optimization problem with a spherical constraint set, and hence is naturally phrased in the language of manifold optimization. To show this apparently hard problem is tractable, we prove that with high probability (w.h.p.) the function lies in the $\mathcal{X}$ family (Chapter 2), particularly devoid of spurious local minimizers. This benign geometric structure allows us to design a Riemannian trust-region algorithm over the sphere that provably converges to a local minimizer with an arbitrary initialization, despite the presence of saddle points.

This part is organized as follows. In Chapter 3 we motivate the dictionary learning problem and overview main ingredients of our nonconvex approach. In Chapter 4 we present our main geometric results that confirm the central nonconvex problem to be solved lie in the $\mathcal{X}$ family. In Chapter 5 we present necessary technical machinery and results for convergence proof of the Riemannian trust-region algorithm over the sphere. Solving the nonconvex problem recovers one row of $\boldsymbol{X}_0$ each time. We will present an algorithmic pipeline that entails solving multiple instances of the nonconvex problem to sequentially recover all rows of $\boldsymbol{X}_0$, and hence also $\boldsymbol{A}_0$, in Chapter 6. After presenting simulations to corroborate our theory in Chapter 7, we wrap up the main content in Chapter 8 by discussing possible improvement and future directions. All major proofs of geometrical and algorithmic results are deferred to Chapter 9 and Chapter 10, respectively. Chapter 11 augments the technical content presented in Chapter 6. Recurring technical tools and auxiliary results for the proofs are included in Appendix A and Appendix B.

This part is based on our technical report:

Complete Dictionary Recovery over the Sphere. `http://arxiv.org/abs/1504.06785`

The codes to reproduce all the figures and experimental results can be found online:

`https://github.com/sunju/dl_focm`

Chapter 3

Introduction

> ...in effect, uncovering the optimal codebook structure of naturally occurring data involves more challenging empirical questions than any that have ever been solved in empirical work in the mathematical sciences.
>
> Donoho et al [DVDD98], in *Data compression and harmonic analysis*

Given p signal samples from $\mathbb{R}^n$, i.e., $\boldsymbol{Y} \doteq [\boldsymbol{y}_1, \dots, \boldsymbol{y}_p] \in \mathbb{R}^{n \times p}$, is it possible to construct an m-element *dictionary* $\boldsymbol{A} \doteq [\boldsymbol{a}_1, \dots, \boldsymbol{a}_m]$ with m much smaller than p, such that $\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X}$ and the *coefficient* matrix $\boldsymbol{X}$ has as few nonzeros as possible? In other words, this model *dictionary learning* (DL) problem seeks a concise representation for a collection of input signals. Concise signal representations play a central role in signal compression, and also prove useful for many other important tasks, such as signal acquisition, denoising, and classification.

Traditionally, concise signal representations have relied heavily on explicit analytic bases constructed in nonlinear approximation and harmonic analysis. This constructive approach has proved highly successfully; the numerous theoretical advances in these fields (see, e.g., [DeV98, Tem03, DeV09, Can02, MP10a] for summary of relevant results) provide ever more powerful representations, ranging from the classic Fourier to modern multidimensional, multidirectional, multiresolution bases, including wavelets, curvelets, ridgelets, and so on. However, two challenges confront practitioners in adapting these results to new domains: which function class best describes signals at hand, and consequently which representation is most appropriate. These challenges are coupled, as function classes with known "good" analytic bases are rare.

Around 1996, neuroscientists Olshausen and Field discovered that sparse coding, the principle of encoding a signal with few atoms from a learned dictionary, reproduces important properties of the receptive fields of the simple cells that perform early visual processing [OF96, OF97]. The discovery has spurred a flurry of algorithmic developments and successful applications for DL in the past two decades, spanning classical image processing, visual recognition, compressive signal acquisition, and also recent deep architectures for signal classification (see, e.g., [Ela10, MBP14] for review this development).

The learning approach is particularly relevant to modern signal processing and machine learning, which deal with data of huge volume and great variety (e.g., images, audios, graphs, texts, genome sequences, time series, etc). The proliferation of problems and data seems to preclude analytically deriving optimal representations for each new class of data in a timely manner. On the other hand, as datasets grow, learning dictionaries directly from data looks increasingly attractive and promising. When armed with sufficiently many data samples of one signal class, by solving the model DL problem, one would expect to obtain a dictionary that allows sparse representation for the whole class. This hope has been borne out in a number of successful examples [Ela10, MBP14] and theories [MP10b, VMB11, MG13, GJB^{+}13].

3.1 Theoretical and algorithmic challenges

In contrast to the above empirical successes, the theoretical study of DL is still developing. For applications in which DL is to be applied in a "hands-free" manner, it is desirable to have efficient algorithms which are guaranteed to perform correctly, when the input data admit a sparse model. There have been several important recent results in this direction, which we will review in Section 3.4, after our sketching main results. Nevertheless, obtaining algorithms that provably succeed under broad and realistic conditions remains an important research challenge.

To understand where the difficulties arise, we can consider a model formulation, in which we attempt to obtain the dictionary $\boldsymbol{A}$ and coefficients $\boldsymbol{X}$ which best trade-off sparsity and fidelity to the observed data:

$$\mathrm{minimize}_{\boldsymbol{A}\in\mathbb{R}^{n\times m},\boldsymbol{X}\in\mathbb{R}^{m\times p}}\ \ \lambda\left\|\boldsymbol{X}\right\|_1+\frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{X}-\boldsymbol{Y}\right\|_F^2,\ \text{subject to}\ \ \boldsymbol{A}\in\mathcal{A}. \tag{3.1.1}$$

Here, $\left\|\boldsymbol{X}\right\|_1 \doteq \sum_{i,j}\left|X_{ij}\right|$ promotes sparsity of the coefficients, $\lambda \geq 0$ trades off the level of coefficient sparsity and quality of approximation, and $\mathcal{A}$ imposes desired structures on the dictionary.

This formulation is nonconvex: the admissible set $\mathcal{A}$ is typically nonconvex (e.g., orthogonal group, matrices with normalized columns)[1], while the most daunting nonconvexity comes from the bilinear mapping:

[1] For example, in nonlinear approximation and harmonic analysis, orthonormal basis or (tight-)frames are preferred; to fix the scale

$(\boldsymbol{A}, \boldsymbol{X}) \mapsto \boldsymbol{A}\boldsymbol{X}$. Because $(\boldsymbol{A}, \boldsymbol{X})$ and $\left(\boldsymbol{A}\boldsymbol{\Pi}\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\Pi}^*\boldsymbol{X}\right)$ result in the same objective value for the model formulation (3.1.1), where $\boldsymbol{\Pi}$ is any permutation matrix, and $\boldsymbol{\Sigma}$ any diagonal matrix with $\{\pm 1\}$ on its diagonal, we should expect the problem to have combinatorially many global minimizers. Moreover, these minimizers are generally isolated, and hence the problem does not appear to be amenable to convex relaxation (see similar discussions in, e.g., [GS10] and [GW11]).[2] This contrasts sharply with problems in sparse recovery and compressed sensing, in which simple convex relaxations are often provably effective [DT09, OH10, CLMW11, DGM13, MT14, MHWG14, CRPW12, CSV13, ALMT14, Can14]. Is there any hope to obtain global solutions to the DL problem? The numerical surprise we encountered at the start of this thesis (i.e., Section 1.1) supports a positive answer.

3.2 Dictionary recovery and our results

In this part (Part II), we take a step towards explaining the surprising effectiveness of simple optimization methods for DL. We focus on the *dictionary recovery* (DR) setting: given a data matrix $\boldsymbol{Y}$ generated as $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$, where $\boldsymbol{A}_0 \in \mathcal{A} \subseteq \mathbb{R}^{n \times m}$ and $\boldsymbol{X}_0 \in \mathbb{R}^{m \times p}$ is "reasonably sparse", try to recover $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$. Here recovery means to return any pair $\left(\boldsymbol{A}_0\boldsymbol{\Pi}\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\Pi}^*\boldsymbol{X}_0\right)$, where $\boldsymbol{\Pi}$ is a permutation matrix and $\boldsymbol{\Sigma}$ is a nonsingular diagonal matrix, i.e., recovering up to sign, scale, and permutation.

To define a reasonably simple and structured problem, we make the following assumptions:

- The target dictionary $\boldsymbol{A}_0$ is complete, i.e., square and invertible ($m = n$). In particular, this class includes orthogonal dictionaries. Admittedly overcomplete dictionaries tend to be more powerful for modeling and to allow sparser representations. Nevertheless, most classic hand-designed dictionaries in common use are orthogonal. Orthobases are competitive in performance for certain tasks such as image denoising [BCJ13], and admit faster algorithms for learning and encoding. [3]

- The coefficient matrix $\boldsymbol{X}_0$ follows the Bernoulli-Gaussian (BG) model with rate θ: $[\boldsymbol{X}_0]_{ij} = \Omega_{ij} V_{ij}$, with

ambiguity discussed in the text, a common practice is to require that $\boldsymbol{A}$ to be column-normalized. There is no obvious reason to believe that convexifying these constraint sets would leave the optima unchanged. For example, the convex hull of the orthogonal group O_n is the operator norm ball $\left\{\boldsymbol{X} \in \mathbb{R}^{n \times n} : \|\boldsymbol{X}\| \le 1\right\}$. If there are no effective symmetry breaking constraints, any convex objective function tends to have minimizers inside the ball, which obviously will not be orthogonal matrices. Other ideas such as lifting may not play together with the objective function, nor yield tight relaxations (see, e.g., [BKS13a, BR14]).

[2] Semidefinite programming (SDP) lifting may be one useful general strategy to convexify bilinear inverse problems, see, e.g., [ARR14, CM14]. However, for problems with general nonlinear constraints, it is unclear whether the lifting always yield tight relaxation, consider, e.g., [BKS13a, BR14] again.

[3] Empirically, there is no systematic evidence supporting that overcomplete dictionaries are strictly necessary for good performance in all published applications (though [OF97] argues for the necessity from neuroscience perspective). Some of the ideas and tools developed here for complete dictionaries may also apply to certain classes of structured overcomplete dictionaries, such as tight frames.

$\Omega_{ij} \sim \mathrm{Ber}(\theta)$ and $V_{ij} \sim \mathcal{N}(0,1)$, where all the different random variables are jointly independent. We write compactly $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}(\theta)$.

We prove the following result:

> **Theorem 3.1 (Informal statement of our results)** *For any $\theta \in (0, 1/3)$, given $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$ with $\boldsymbol{A}_0$ a complete dictionary and $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}(\theta)$, there is a polynomial-time algorithm that recovers $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ with high probability (at least $1 - O(p^{-6})$) whenever $p \geq p_\star(n, 1/\theta, \kappa(\boldsymbol{A}_0), 1/\mu)$ for a fixed polynomial $p_\star(\cdot)$, where $\kappa(\boldsymbol{A}_0)$ is the condition number of $\boldsymbol{A}_0$ and μ is a parameter that can be set as $cn^{-5/4}$ for a fixed constant $c > 0$.*

Obviously, even if $\boldsymbol{X}_0$ is known, one needs $p \geq n$ to make the identification problem well posed. Under our particular probabilistic model, a simple coupon collection argument implies that one needs $p \geq \Omega\left(\frac{1}{\theta}\log n\right)$ to ensure all atoms in $\boldsymbol{A}_0$ are observed with high probability (w.h.p.). To ensure that an efficient algorithm exists may demand more. Our result implies when p is polynomial in n, $1/\theta$ and $\kappa(\boldsymbol{A}_0)$, recovery with efficient algorithm is possible.

The parameter θ controls the sparsity level of $\boldsymbol{X}_0$. Intuitively, the recovery problem is easy for small θ and becomes harder for large θ.[4] It is perhaps surprising that an efficient algorithm can succeed up to constant θ, i.e., linear sparsity in $\boldsymbol{X}_0$. Compared to the case when $\boldsymbol{A}_0$ is known, there is only at most a constant gap in the sparsity level one can deal with.

For DR, our result gives the first efficient algorithm that provably recovers complete $\boldsymbol{A}_0$ and sparse $\boldsymbol{X}_0$ when $\boldsymbol{X}_0$ has $O(n)$ nonzeros per column under appropriate probability model. Section 3.4 provides detailed comparison of our result with other recent recovery results for complete and overcomplete dictionaries.

3.3 Main ingredients and innovations

In this section we describe three main ingredients that we use to obtain the stated result.

3.3.1 A nonconvex formulation

Since $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$ and $\boldsymbol{A}_0$ is complete, $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)$ ($\mathrm{row}(\cdot)$ denotes the row space of a matrix) and hence rows of $\boldsymbol{X}_0$ are sparse vectors in the known (linear) subspace $\mathrm{row}(\boldsymbol{Y})$. We can use this fact to first recover the rows of $\boldsymbol{X}_0$, and subsequently recover $\boldsymbol{A}_0$ by solving a system of linear equations. In

[4] Indeed, when θ is small enough such that columns of $\boldsymbol{X}_0$ are predominately 1-sparse, one directly observes scaled versions of the atoms (i.e., columns of $\boldsymbol{X}_0$); when $\boldsymbol{X}_0$ is fully dense corresponding to $\theta = 1$, recovery is never possible as one can easily find another complete $\boldsymbol{A}_0'$ and fully dense $\boldsymbol{X}_0'$ such that $\boldsymbol{Y} = \boldsymbol{A}_0'\boldsymbol{X}_0'$ with $\boldsymbol{A}_0'$ not equivalent to $\boldsymbol{A}_0$.

fact, for $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}\left(\theta\right)$, rows of $\boldsymbol{X}_0$ are the n *sparsest* vectors (directions) in $\mathrm{row}\left(\boldsymbol{Y}\right)$ w.h.p. whenever $p \geq \Omega\left(n \log n\right)$ [SWW12]. Thus one might try to recover rows of $\boldsymbol{X}_0$ by solving

$$\text{minimize } \left\|\boldsymbol{q}^* \boldsymbol{Y}\right\|_0 \text{ subject to } \boldsymbol{q} \neq \boldsymbol{0}. \tag{3.3.1}$$

The objective is discontinuous, and the domain is an open set. In particular, the homogeneity constraint is unconventional and tricky to deal with. Since the recovery is up to scale, one can remove the homogeneity by fixing the scale of $\boldsymbol{q}$. Known relaxations [SWW12, DH14] fix the scale by setting $\left\|\boldsymbol{q}^* \boldsymbol{Y}\right\|_\infty = 1$, where $\left\|\cdot\right\|_\infty$ is the elementwise ℓ^∞ norm. The optimization problem reduces to a sequence of convex programs, which recover $\left(\boldsymbol{A}_0, \boldsymbol{X}_0\right)$ for very sparse $\boldsymbol{X}_0$, but provably break down when columns of $\boldsymbol{X}_0$ has more than $O\left(\sqrt{n}\right)$ nonzeros, or $\theta \geq \Omega\left(1/\sqrt{n}\right)$. Inspired by the success of nonconvex heuristic in our image experiment (Section 1.1), we work with a *nonconvex* alternative[5]:

$$\text{minimize } f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \doteq \frac{1}{p} \sum_{k=1}^{p} h_\mu\left(\boldsymbol{q}^* \widehat{\boldsymbol{y}}_k\right), \text{ subject to } \left\|\boldsymbol{q}\right\| = 1, \tag{3.3.2}$$

where $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{n \times p}$ is a proxy for $\boldsymbol{Y}$ (i.e., after appropriate processing), k indexes columns of $\widehat{\boldsymbol{Y}}$, and $\left\|\cdot\right\|$ is the usual ℓ^2 norm for vectors. Here $h_\mu\left(\cdot\right)$ is chosen to be a convex smooth approximation to $\left|\cdot\right|$, as plotted against the $\left|\cdot\right|$ function in Figure 3.1; namely,

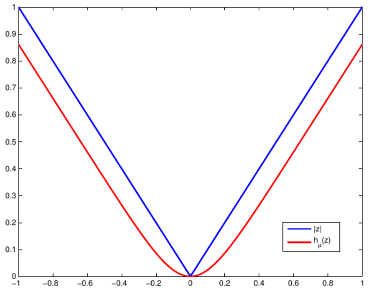


Figure 3.1: Illustration of the smooth ℓ^1 surrogate used in sparse dictionary recovery. The function is chosen to be infinitely differentiable such that later on second-order information can be extracted from second-order derivatives.

$$h_\mu\left(z\right) = \mu \log \cosh(z/\mu), \tag{3.3.3}$$

which is infinitely differentiable and μ controls the smoothing level.[6] The spherical constraint is nonconvex. Hence, a-priori, it is unclear whether (3.3.2) admits efficient algorithms that attain global optimizers. Surprisingly, simple descent algorithms for (3.3.2) exhibit very striking behavior: on many practical numerical

[5] A similar formulation was proposed in [ZP01] in the context of blind source separation; see also [QSW14].

[6] In fact, there is nothing special about this choice and we believe that any valid smooth (twice continuously differentiable) approximation to $\left|\cdot\right|$ would work and yield qualitatively similar results. We also have some preliminary results showing the latter geometric picture remains the same for certain nonsmooth functions, such as a modified version of the Huber function, though the analysis involves handling a different set of technical subtleties. The algorithm also needs additional modifications.

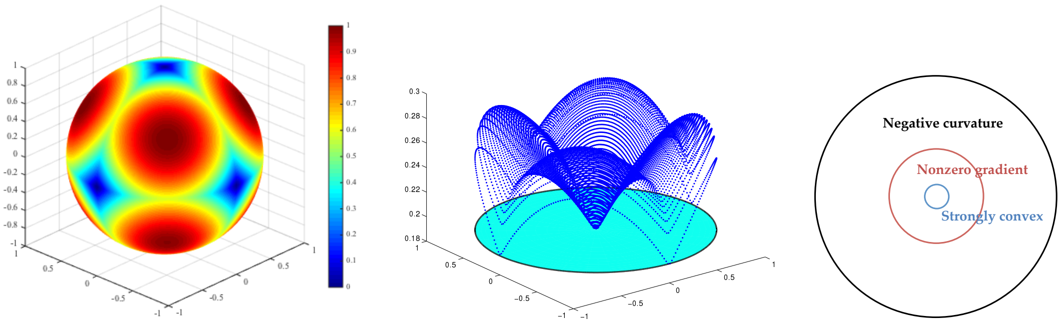


Figure 3.2: Why is dictionary learning over $\mathbb{S}^{n-1}$ tractable? Assume the target dictionary $\boldsymbol{A}_0$ is orthogonal. **Left:** Large sample objective function $\mathbb{E}_{\boldsymbol{X}_0}\left[f\left(\boldsymbol{q}\right)\right]$. The only local minimizers are the columns of $\boldsymbol{A}_0$ and their negatives. **Center:** the same function, visualized as a height above the plane $\boldsymbol{a}_1^{\perp}$ ($\boldsymbol{a}_1$ is the first column of $\boldsymbol{A}_0$). **Right:** Around the optimizer, the function exhibits a small region of strong convexity, a region of strong gradient, and finally a region in which the direction away from the target minimizer is a direction of negative curvature.

examples[7], they appear to produce global solutions. Our next section will uncover interesting geometrical structures underlying the phenomenon.

3.3.2 A glimpse into high-dimensional function landscape

For the moment, suppose $\boldsymbol{A}_0$ is orthogonal, and take $\widehat{\boldsymbol{Y}} = \boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$ in (3.3.2). Figure 3.2 (left) plots $\mathbb{E}_{\boldsymbol{X}_0}\left[f\left(\boldsymbol{q};\boldsymbol{Y}\right)\right]$ over $\mathbb{S}^2$ ($n = 3$). Remarkably, $\mathbb{E}_{\boldsymbol{X}_0}\left[f\left(\boldsymbol{q};\boldsymbol{Y}\right)\right]$ has no spurious local minimizers. In fact, every local minimizer $\widehat{\boldsymbol{q}}$ produces a row of $\boldsymbol{X}_0$: $\widehat{\boldsymbol{q}}^*\boldsymbol{Y} = \alpha\boldsymbol{e}_i^*\boldsymbol{X}_0$ for some $\alpha \neq 0$. Moreover, there are saddle points, each with an obvious negative curvature direction connecting two neighboring local minimizers. Thus, qualitatively $\mathbb{E}_{\boldsymbol{X}_0}\left[f\left(\boldsymbol{q};\boldsymbol{Y}\right)\right]$ is an $\mathcal{X}$ function over $\mathbb{S}^{n-1}$.

To better illustrate the point, we take the particular case $\boldsymbol{A}_0 = \boldsymbol{I}$ and project the upper hemisphere above the equatorial plane $\boldsymbol{e}_3^{\perp}$ onto $\boldsymbol{e}_3^{\perp}$. The projection is bijective and we equivalently define a reparameterization $g : \boldsymbol{e}_3^{\perp} \mapsto \mathbb{R}$ of f. Figure 3.2 (center) plots the graph of g. Obviously the only local minimizers are $\boldsymbol{0}, \pm\boldsymbol{e}_1, \pm\boldsymbol{e}_2$, and they are also global minimizers. Moreover, the apparent nonconvex landscape has interesting structures around $\boldsymbol{0}$: when moving away from $\boldsymbol{0}$, one sees successively a strongly convex region, a nonzero gradient region, and a region where at each point one can always find a direction of negative curvature, as shown schematically in Figure 3.2 (right). This geometry implies that at any nonoptimal point, there is always at least one direction of descent. Thus, any algorithm that can take advantage of the descent directions will likely converge to one global minimizer, irrespective of initialization.

Two challenges stand out when implementing this idea. For geometry, one has to show similar structure

[7] ... not restricted to the model we assume here for $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$.

exists for general complete $\boldsymbol{A}_0$, in high dimensions ($n \geq 3$), when the number of observations p is finite (vs. the expectation in the experiment). For algorithms, we need to be able to take advantage of this structure without knowing $\boldsymbol{A}_0$ ahead of time. In Section 3.3.3, we describe a trust-region method over the sphere which addresses the latter challenge.

Geometry for orthogonal $\boldsymbol{A}_0$. In this case, we take $\widehat{\boldsymbol{Y}} = \boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$. Since $f\left(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0\right) = f\left(\boldsymbol{A}_0^*\boldsymbol{q}; \boldsymbol{X}_0\right)$, the landscape of $f\left(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0\right)$ is simply a rotated version of that of $f\left(\boldsymbol{q}; \boldsymbol{X}_0\right)$, i.e., when $\boldsymbol{A}_0 = \boldsymbol{I}$. Hence we will focus on the case when $\boldsymbol{A}_0 = \boldsymbol{I}$. Among the $2n$ symmetric sections of $\mathbb{S}^{n-1}$ centered around the signed basis vectors $\pm\boldsymbol{e}_1, \dots, \pm\boldsymbol{e}_n$, we work with the symmetric section around $\boldsymbol{e}_n$ as an example. The result will carry over to all sections with analogous arguments; together this provides a complete characterization of the function $f\left(\boldsymbol{q}; \boldsymbol{X}_0\right)$ over $\mathbb{S}^{n-1}$.

We again invoke the projection trick described above, this time onto the equatorial plane $\boldsymbol{e}_n^\perp$. This can be formally captured by the reparameterization mapping:

$$\boldsymbol{q}\left(\boldsymbol{w}\right) = \left(\boldsymbol{w}, \sqrt{1 - \left\|\boldsymbol{w}\right\|^2}\right), \ \boldsymbol{w} \in \mathbb{B}^{n-1}, \tag{3.3.4}$$

where $\boldsymbol{w}$ is the new variable in $\boldsymbol{e}_n^\perp \cap \mathbb{B}^{n-1}$ and $\mathbb{B}^{n-1}$ is the unit ball in $\mathbb{R}^{n-1}$. We first study the composition $g\left(\boldsymbol{w}; \boldsymbol{X}_0\right) \doteq f\left(\boldsymbol{q}\left(\boldsymbol{w}\right); \boldsymbol{X}_0\right)$ over the set

$$\Gamma \doteq \left\{\boldsymbol{w} : \left\|\boldsymbol{w}\right\| < \sqrt{\tfrac{4n-1}{4n}}\right\}. \tag{3.3.5}$$

It can be verified the section we chose to work with is contained in this set[8].

Our analysis characterizes the properties of $g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)$ by studying three quantities

$$\nabla^2 g\left(\boldsymbol{w}; \boldsymbol{X}_0\right), \quad \frac{\boldsymbol{w}^*\nabla g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)}{\left\|\boldsymbol{w}\right\|}, \quad \frac{\boldsymbol{w}^*\nabla^2 g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)\boldsymbol{w}}{\left\|\boldsymbol{w}\right\|^2}$$

respectively over three consecutive regions moving away from the origin, corresponding to the three regions in Figure 3.2 (right). In particular, through typical expectation-concentration style argument, we show that (Theorem 4.1 & Corollary 4.2) there exists a positive constant c such that

$$\nabla^2 g\left(\boldsymbol{w}; \boldsymbol{X}_0\right) \succeq \frac{1}{\mu}c\theta\boldsymbol{I}, \quad \frac{\boldsymbol{w}^*\nabla g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)}{\left\|\boldsymbol{w}\right\|} \geq c\theta, \quad \frac{\boldsymbol{w}^*\nabla^2 g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)\boldsymbol{w}}{\left\|\boldsymbol{w}\right\|^2} \leq -c\theta \tag{3.3.6}$$

over the respective regions w.h.p., confirming our low-dimensional observations described above. In particu-

[8] Indeed, if $\langle \boldsymbol{q}, \boldsymbol{e}_n\rangle \geq |\langle \boldsymbol{q}, \boldsymbol{e}_i\rangle|$ for any $i \neq n$, $1 - \left\|\boldsymbol{w}\right\|^2 = q_n^2 \geq 1/n$, implying $\left\|\boldsymbol{w}\right\|^2 \leq \frac{n-1}{n} < \frac{4n-1}{4n}$. The reason we have defined an open set instead of a closed (compact) one is to avoid potential "artificial" local minimizers located on the boundary.

lar, the favorable structure we observed for $n = 3$ persists in high dimensions, w.h.p., even when p is large *yet finite*, for the case $\boldsymbol{A}_0$ is orthogonal. Moreover, the local minimizer of $g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)$ over Γ is very close to $\mathbf{0}$, within a distance of $O\left(\mu\right)$.

Geometry for complete $\boldsymbol{A}_0$. For general complete dictionaries $\boldsymbol{A}_0$, we hope that the function f retains the nice geometric structure discussed above. We can ensure this by "preconditioning" $\boldsymbol{Y}$ such that the output looks as if being generated from a certain orthogonal matrix, possibly plus a small perturbation. We can then argue that the perturbation does not significantly affect the properties of the graph of the objective function. Write

$$\overline{\boldsymbol{Y}} = \sqrt{p\theta}\left(\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}. \tag{3.3.7}$$

Note that for $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}\left(\theta\right)$, $\mathbb{E}\left[\boldsymbol{X}_0\boldsymbol{X}_0^*\right]/\left(p\theta\right) = \boldsymbol{I}$. Thus, one expects $\boldsymbol{Y}\boldsymbol{Y}^*/\left(p\theta\right) = \boldsymbol{A}_0\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{A}_0^*/\left(p\theta\right)$ to behave roughly like $\boldsymbol{A}_0\boldsymbol{A}_0^*$ and hence $\overline{\boldsymbol{Y}}$ to behave like

$$\left(\boldsymbol{A}_0\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 \tag{3.3.8}$$

where we write the SVD of $\boldsymbol{A}_0$ as $\boldsymbol{A}_0 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. It is easy to see $\boldsymbol{U}\boldsymbol{V}^*$ is an orthogonal matrix. Hence the preconditioning scheme we have introduced is technically sound.

Our analysis shows that $\overline{\boldsymbol{Y}}$ can be written as

$$\overline{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \boldsymbol{\Xi}\boldsymbol{X}_0, \tag{3.3.9}$$

where $\boldsymbol{\Xi}$ is a matrix with small magnitude. Simple perturbation argument shows that (Theorem 4.3 & Corollary 4.4) the constant c in (3.3.6) is at most shrunk to $c/2$ for all $\boldsymbol{w}$ when p is sufficiently large. Thus, the qualitative aspects of the geometry have not been changed by the perturbation.

Our $\boldsymbol{w}$ space calculation confirms that $g(\boldsymbol{w})$ is an $\mathcal{X}$ function with concrete parameters. We will discuss the implication on the original function $f(\boldsymbol{q})$ after our main geometric results in Section 4.1.

3.3.3 A second-order algorithm on the sphere: Riemannian trust-region method

We do not know $\boldsymbol{A}_0$ ahead of time, so our algorithm needs to take advantage of the structure described above without knowledge of $\boldsymbol{A}_0$. Intuitively, this seems possible as the descent direction in the $\boldsymbol{w}$ space appears to also be a local descent direction for f over the sphere. Another issue is that although the optimization problem has no spurious local minimizers, it does have many ridable saddle points (Figure 3.2). As discussed in Section 2.4, we can use second-order information to guarantee to escape from ridable saddle points in a trust-region framework. We specialize the Riemannian trust-region method in Section 2.4 to the sphere.

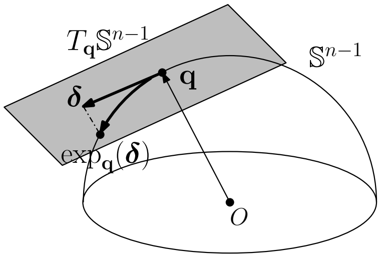

$T_{\mathbf{q}}\mathbb{S}^{n-1}$
$\mathbb{S}^{n-1}$
$\mathbf{q}$
$\boldsymbol{\delta}$
$\exp_{\mathbf{q}}(\boldsymbol{\delta})$
O

Figure 3.3: Illustrations of the tangent space $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ and exponential map $\exp_{\boldsymbol{q}}(\boldsymbol{\delta})$ defined on the sphere $\mathbb{S}^{n-1}$.

Consider an iterate sequence $\boldsymbol{q}^{(0)}, \boldsymbol{q}^{(1)}, \boldsymbol{q}^{(2)}, \dots$ over $\mathbb{S}^{n-1}$. At any $\boldsymbol{q} \in \mathbb{S}^{n-1}$, the tangent space to the sphere is $T_{\boldsymbol{q}}\mathbb{S}^{n-1} \doteq \{\boldsymbol{v} : \boldsymbol{v}^*\boldsymbol{q} = 0\}$ (see Figure 3.3) and the exponential map is

$$\exp_{\boldsymbol{q}}(\boldsymbol{\delta}) = \boldsymbol{q}\cos\|\boldsymbol{\delta}\| + \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}\sin\|\boldsymbol{\delta}\| .$$

Thus, to form a local quadratic approximation to $f(\boldsymbol{q})$ around the current iterate $\boldsymbol{q}^{(k)}$, we consider $f(\exp_{\boldsymbol{q}^{(k)}}(\boldsymbol{\delta})) : T_{\boldsymbol{q}^{(k)}} \mapsto \mathbb{R}$ and its second-order Taylor approximation

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{q}^{(k)}) \ \doteq \ f(\boldsymbol{q}^{(k)}) + \left\langle \nabla f(\boldsymbol{q}^{(k)}), \boldsymbol{\delta} \right\rangle + \frac{1}{2}\boldsymbol{\delta}^* \left(\nabla^2 f(\boldsymbol{q}^{(k)}) - \left\langle \nabla f(\boldsymbol{q}^{(k)}), \boldsymbol{q}^{(k)} \right\rangle \boldsymbol{I} \right) \boldsymbol{\delta}. \tag{3.3.10}$$

Let $\mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}} \doteq \boldsymbol{I} - \boldsymbol{q}^{(k)} \left(\boldsymbol{q}^{(k)}\right)^*$ be the orthoprojector onto $T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}$. The Riemannian gradient and Hessian can be read from the above quadratic approximation as:

$$\begin{aligned} \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right) &\doteq \mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}} \nabla f(\boldsymbol{q}^{(k)}), \\ \operatorname{Hess} f\left(\boldsymbol{q}^{(k)}\right) &\doteq \mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}} \left(\nabla^2 f(\boldsymbol{q}^{(k)}) - \left\langle \nabla f(\boldsymbol{q}^{(k)}), \boldsymbol{q}^{(k)} \right\rangle \boldsymbol{I} \right) \mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}}. \end{aligned}$$

One can then deploy the transformation trick discussed in Section 2.4 to solve the Riemannian trust-region subproblem

$$\operatorname{minimize}_{\boldsymbol{\delta} \in T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1},\ \|\boldsymbol{\delta}\|_2 \le \Delta} \widehat{f}\left(\boldsymbol{\delta}; \boldsymbol{q}^{(k)}\right). \tag{3.3.11}$$

Once an optimizer $\boldsymbol{\delta}_\star$ is obtained, the next iterate is determined as

$$\boldsymbol{q}^{(k+1)} \doteq \exp_{\boldsymbol{q}^{(k)}}(\boldsymbol{\delta}_\star) = \boldsymbol{q}^{(k)}\cos\|\boldsymbol{\delta}_\star\| + \tfrac{\boldsymbol{\delta}_\star}{\|\boldsymbol{\delta}_\star\|}\sin\|\boldsymbol{\delta}_\star\| . \tag{3.3.12}$$

As seen from Figure 3.3, the movement to the next iterate is “along the direction"[9] of $\boldsymbol{\delta}_\star$ while staying over the sphere.

[9] Technically, moving along the geodesic whose velocity at time zero is $\boldsymbol{\delta}_\star$.

Based on the geometric characterizations, we prove that w.h.p., the algorithm converges to a local minimizer when the parameter Δ is sufficiently small. Specifically, we show that (1) a trust-region step in the negative curvature and strong gradient regions decreases the objective value by at least a fixed amount; (2) the trust-region iterate sequence will finally move to and stay in the strongly convex region, and converge to a local minimizer with an asymptotic quadratic rate. In short, the geometric structure implies that from *any initialization*, the iterate sequence converges to a close approximation to the target solution in a polynomial number of steps.

3.4 Prior arts and connections

It is far too ambitious to include here a comprehensive review of the exciting developments of DL algorithms and applications after the pioneer work [OF96]. We refer the reader to Chapter 12 - 15 of the book [Ela10] and the survey paper [MBP14] for summaries of relevant developments in image analysis and visual recognition. In the following, we focus on reviewing recent developments on the theoretical side of DL, and draw connections to problems and techniques that are relevant to the current work.

Theoretical Dictionary Learning. The theoretical study of DL in the recovery setting started only very recently. [AEB06] was the first to provide an algorithmic procedure to correctly extract the generating dictionary. The algorithm requires exponentially many samples and has exponential running time; see also [HS11]. Subsequent work [GS10, GW11, Sch14a, Sch14b, Sch15] studied when the target dictionary is a local optimum of natural recovery criteria ("local correctness"). These meticulous analyses show that polynomially many samples are sufficient to ensure local correctness under natural assumptions. However, these results do not imply that one can design efficient algorithms to obtain the desired local optimizer and hence the dictionary.

[SWW12] initiated the on-going research effort to provide efficient algorithms that globally solve DR. They showed that one can recover a complete dictionary $\boldsymbol{A}_0$ from $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0$ by solving a certain sequence of linear programs, when $\boldsymbol{X}_0$ is a sparse random matrix with $O(\sqrt{n})$ nonzeros per column. [AAJ$^+$13, AAN13] and [AGM13, AGMM15] give efficient algorithms that provably recover overcomplete ($m \geq n$) and incoherent dictionaries, based on a combination of {clustering or spectral initialization} and local refinement. These algorithms again succeed when $\boldsymbol{X}_0$ has $\widetilde{O}(\sqrt{n})$ [10] nonzeros per column. Recent work [BKS14] provides

[10] The $\widetilde{O}$ suppresses some logarithm factors.

the first polynomial-time algorithm that provably recovers most "nice" overcomplete dictionaries when $\boldsymbol{X}_0$ has $O(n^{1-\delta})$ nonzeros per column for any constant $\delta \in (0,1)$. However, the proposed algorithm runs in super-polynomial time when the sparsity level goes up to $O(n)$. Similarly, [ABGM14] also proposes a super-polynomial (quasipolynomial) time algorithm that guarantees recovery with (almost) $O(n)$ nonzeros per column. By comparison, we give the first *polynomial-time* algorithm that provably recovers complete dictionary $\boldsymbol{A}_0$ when $\boldsymbol{X}_0$ has $O(n)$ nonzeros per column.

Aside from efficient recovery, other theoretical work on DL includes results on identifiability [AEB06, HS11, WY15], generalization bounds [MP10b, VMB11, MG13, GJB$^+$13], and noise stability [GJB14].

Finding Sparse Vectors in a Linear Subspace. We have followed [SWW12] and cast the core problem as finding the sparsest vectors in a given linear subspace, which is also of independent interest. Under a planted sparse model[11], [DH14] shows solving a sequence of linear programs similar to [SWW12] can recover sparse vectors with sparsity up to $O(p/\sqrt{n})$, sublinear in the vector dimension. [QSW14] improved the recovery limit to $O(p)$ by solving a nonconvex sphere constrained problem similar to (3.3.2)[12] via an alternating direction algorithm. The idea of seeking rows of $\boldsymbol{X}_0$ sequentially by solving the above core problem sees precursors in [ZP01] for blind source separation, and [GN10] for matrix sparsification. [ZP01] also proposed a nonconvex optimization similar to (3.3.2) here and that employed in [QSW14].

Nonconvex Optimization Problems. For other nonconvex optimization problems of recovery of structured signals[13], including low-rank matrix completion/recovery [KMO10, JNS13, Har14, HW14, NNS$^+$14, JN14, SL14, ZL15, TBSR15, CW15], phase retreival [NJS13, CLS15b, CC15, WWS15], tensor recovery [JO14, AGJ14b, AGJ14a, AJSN15], mixed regression [YCS13, LWB13], structured element pursuit [QSW14], and recovery of simultaneously structured signals [LWB13], numerical linear algebra [JJKN15], the initialization plus local refinement strategy adopted in theoretical DL [AAJ$^+$13, AAN13, AGM13, AGMM15, ABGM14] is also crucial: nearness to the target solution enables exploiting the local geometry of the target to analyze the local refinement.[14] By comparison, we provide a complete characterization of the global geometry, which admits efficient algorithms without any special initialization. The idea of separating the geometric analysis and

[11] ... where one sparse vector embedded in an otherwise random subspace.

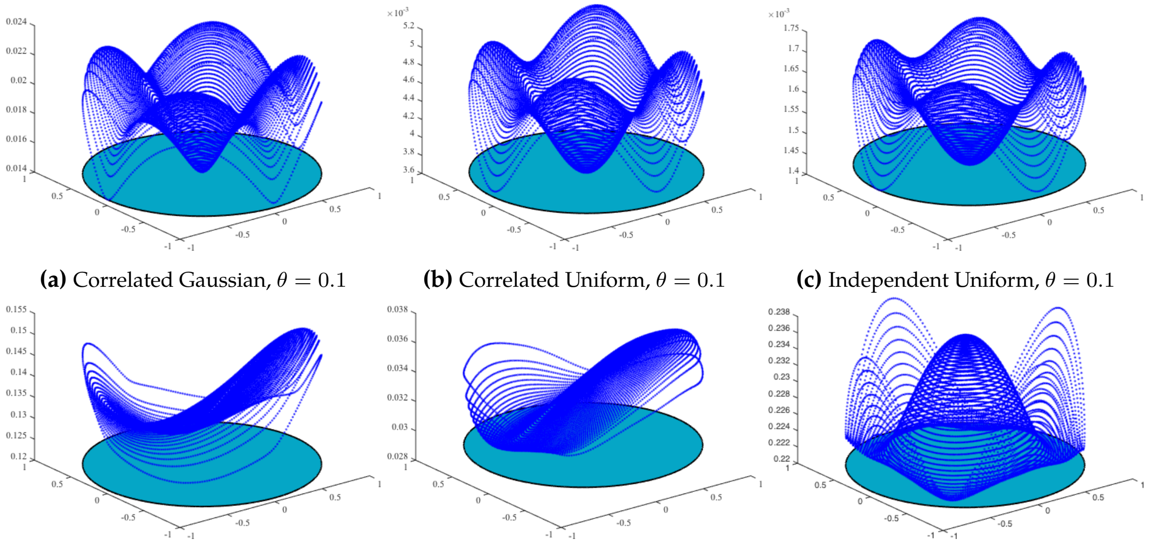
[12] The only difference is that they chose to work with the Huber function as a proxy of the $\|\cdot\|_1$ function.

[13] This is a body of recent work studying nonconvex recovery up to statistical precision, including, e.g., [LW11, LW13, WLL14, BWY14, WGNL14, LW14, Loh15, SLLC15].

[14] The powerful framework [ABRS10, BST14] to establish local convergence of ADM algorithms to critical points applies to DL/DR also, see, e.g., [BJQS14, BQJ14, BJS14]. However, these results do not guarantee to produce global optima.

algorithmic design may also prove valuable for other nonconvex problems discussed above.

Optimization over Riemannian Manifolds. Our trust-region algorithm on the sphere builds on the extensive research efforts to generalize Euclidean numerical algorithms to (Riemannian) manifold settings. We refer the reader to the monographs [Udr94, HMG94, AMS09] for survey of developments in this field. In particular, [EAS98] developed Newton and conjugate-gradient methods for the Stiefel manifolds, of which the sphere is a special case. [ABG07] generalized the trust-region methods to Riemannian manifolds. We cannot, however, adopt the existing convergence results that concern either global convergence (convergence to critical points) or local convergence (convergence to a local minimum within a radius), or the forthcoming generic results on convergence to second-order necessary points [BAC16] under weaker assumptions. The particular geometric structure forces us to piece together different arguments to obtain the specialized global result.



(a) Correlated Gaussian, $\theta = 0.1$ **(b)** Correlated Uniform, $\theta = 0.1$ **(c)** Independent Uniform, $\theta = 0.1$

(d) Correlated Gaussian, $\theta = 0.9$ **(e)** Correlated Uniform, $\theta = 0.9$ **(f)** Independent Uniform, $\theta = 1$

Figure 3.4: Asymptotic function landscapes in $\mathbb{R}^3$ when rows of $\boldsymbol{X}_0$ are not independent for sparse dictionary learning. W.l.o.g., we again assume $\boldsymbol{A}_0 = \boldsymbol{I}$. In (a) and (d), $\boldsymbol{X}_0 = \boldsymbol{\Omega} \odot \boldsymbol{V}$, with $\boldsymbol{\Omega} \sim_{i.i.d.} \mathrm{Ber}(\theta)$ and columns of $\boldsymbol{X}_0$ i.i.d. Gaussian vectors obeying $\boldsymbol{v}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^2)$ for symmetric $\boldsymbol{\Sigma}$ with 1's on the diagonal and i.i.d. off-diagonal entries distributed as $\mathcal{N}(0, \sqrt{2}/20)$. Similarly, in (b) and (e), $\boldsymbol{X}_0 = \boldsymbol{\Omega} \odot \boldsymbol{W}$, with $\boldsymbol{\Omega} \sim_{i.i.d.} \mathrm{Ber}(\theta)$ and columns of $\boldsymbol{X}_0$ i.i.d. vectors generated as $\boldsymbol{w}_i = \boldsymbol{\Sigma} \boldsymbol{u}^i$ with $\boldsymbol{u}_i \sim_{i.i.d.} \mathrm{Uniform}[-0.5, 0.5]$. For comparison, in (c) and (f), $\boldsymbol{X}_0 = \boldsymbol{\Omega} \odot \boldsymbol{W}$ with $\boldsymbol{\Omega} \sim_{i.i.d.} \mathrm{Ber}(\theta)$ and $\boldsymbol{W} \sim_{i.i.d.} \mathrm{Uniform}[-0.5, 0.5]$. Here $\odot$ denote the elementwise product, and the objective function is still based on the sparsity surrogate in (3.3.2).

Independent Component Analysis (ICA) and Other Matrix Factorization Problems. DL can also be considered in the general framework of matrix factorization problems, which encompass the classic principal component analysis (PCA), ICA, and clustering, and more recent problems such as nonnegative matrix factorization (NMF), multi-layer neural nets (deep learning architectures). Most of these problems are NP-hard. Identifying tractable cases of practical interest and providing provable efficient algorithms are subject of on-going research endeavors; see, e.g., recent progresses on NMF [AGKM12], and learning deep neural nets [ABGM13, SA14c, NP13, LSSS14].

ICA factors a data matrix $\boldsymbol{Y}$ as $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$ such that $\boldsymbol{A}$ is square and rows of $\boldsymbol{X}$ are as independent as possible [HO00, HKO01]. In theoretical study of the recovery problem, it is often assumed that rows of $\boldsymbol{X}_0$ are (weakly) independent (see, e.g., [Com94, FJK96, AGMS12]). Our i.i.d. probability model on $\boldsymbol{X}_0$ implies rows of $\boldsymbol{X}_0$ are independent, aligning our problem perfectly with the ICA problem. More interestingly, the $\log\cosh$ objective we analyze here was proposed as a general-purpose *contrast function* in ICA that has not been thoroughly analyzed [Hyv99], and algorithm and analysis with another popular contrast function, the fourth-order cumulants, indeed overlap with ours considerably [FJK96, AGMS12][15]. While this interesting connection potentially helps port our analysis to ICA, it is a fundamental question to ask what is playing the vital role for DR, sparsity or independence.

Figure 3.4 helps shed some light in this direction, where we again plot the asymptotic objective landscape with the natural reparameterization as in Section 3.3.2. From the left and central panels, it is evident even without independence, $\boldsymbol{X}_0$ with sparse columns induces the familiar geometric structures we saw in Figure 3.2; such structures are broken when the sparsity level becomes large. We believe all our later analyses can be generalized to the correlated cases we experimented with. On the other hand, from the right panel[16], it seems with independence, the function landscape undergoes a transition as sparsity level grows - target solution goes from minimizers of the objective to the maximizers of the objective. Without adequate knowledge of the true sparsity, it is unclear whether one would like to minimize or maximize the objective. This suggests sparsity, instead of independence, is critical to success of our method for DR.

[15] Nevertheless, the objective functions are apparently different. Moreover, we have provided a complete geometric characterization of the objective, in contrast to [FJK96, AGMS12]. We believe the geometric characterization could not only provide insight to the algorithm, but also help improve the algorithm in terms of stability and also finding all components.

[16] We have not showed the results on the BG model here, as it seems the structure persists even when θ approaches 1. We suspect the "phase transition" of the landscape occurs at different points for different distributions and Gaussian is the outlying case where the transition occurs at 1.

Chapter 4

High-Dimensional Function Landscapes

> There is no Royal Road to geometry.
>
> Euclid

In this chapter, we present a quantitative characterization of the objective $f(\boldsymbol{q})$ over the sphere, enriching the qualitative description we provided in Section 3.3.2. To characterize the function landscape of $f(\boldsymbol{q}; \boldsymbol{X}_0)$ over $\mathbb{S}^{n-1}$, we mostly work with the function

$$g(\boldsymbol{w}) \doteq f(\boldsymbol{q}(\boldsymbol{w}); \boldsymbol{X}_0) = \frac{1}{p}\sum_{k=1}^{p} h_\mu\left(\boldsymbol{q}(\boldsymbol{w})^* (\boldsymbol{x}_0)_k\right), \tag{4.0.1}$$

induced by the reparametrization

$$\boldsymbol{q}(\boldsymbol{w}) = \left(\boldsymbol{w}, \sqrt{1 - \|\boldsymbol{w}\|^2}\right), \quad \boldsymbol{w} \in \mathbb{B}^{n-1}. \tag{4.0.2}$$

In particular, we focus our attention to the smaller set

$$\Gamma = \left\{\boldsymbol{w} : \|\boldsymbol{w}\| < \sqrt{\frac{4n-1}{4n}}\right\}, \tag{4.0.3}$$

because $\boldsymbol{q}(\Gamma)$ contains all points $\boldsymbol{q} \in \mathbb{S}^{n-1}$ with $n \in \arg\max_{i \in \pm[n]} \boldsymbol{q}^* \boldsymbol{e}_i$ and we can characterize other parts of f on $\mathbb{S}^{n-1}$ using projection onto other equatorial planes. Note that over Γ, $q_n = \left(1 - \|\boldsymbol{w}\|^2\right)^{1/2} \geq \frac{1}{2\sqrt{n}}$.

Section 4.1 contains precise statements of the geometric results, and discussion of their implications. Section 4.2 collects key intermediate results towards proving the results for orthogonal $\boldsymbol{A}_0$, and Section 4.3 discusses how to extend the results of the orthogonal case to the complete case by a perturbation argument. Detailed proofs to most technical claims are deferred to Chapter 9.

4.1 Main geometric theorems

> **Theorem 4.1 (High-dimensional landscape - orthogonal dictionary)** *Suppose* $\boldsymbol{A}_0 = \boldsymbol{I}$ *and hence* $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0 = \boldsymbol{X}_0$. *There exist positive absolute constants* $c_\star$ *and* C, *such that for any* $\theta \in (0, 1/2)$ *and* $\mu < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}$, *whenever*
>
> $$p \geq \frac{C}{\mu^2\theta^2} n^3 \log \frac{n}{\mu\theta}, \tag{4.1.1}$$
>
> *the following hold simultaneously with high probability:*
>
> $$\nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0) \succeq \frac{c_\star \theta}{\mu} \boldsymbol{I} \qquad \forall\, \boldsymbol{w} \quad s.t. \quad \|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}, \tag{4.1.2}$$
>
> $$\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w}; \boldsymbol{X}_0)}{\|\boldsymbol{w}\|} \geq c_\star \theta \qquad \forall\, \boldsymbol{w} \quad s.t. \quad \frac{\mu}{4\sqrt{2}} \leq \|\boldsymbol{w}\| \leq \frac{1}{20\sqrt{5}} \tag{4.1.3}$$
>
> $$\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} \leq -c_\star \theta \qquad \forall\, \boldsymbol{w} \quad s.t. \quad \frac{1}{20\sqrt{5}} \leq \|\boldsymbol{w}\| \leq \sqrt{\frac{4n-1}{4n}}, \tag{4.1.4}$$
>
> and *the function* $g(\boldsymbol{w}; \boldsymbol{X}_0)$ *has exactly one local minimizer* $\boldsymbol{w}_\star$ *over the open set* $\Gamma \doteq \left\{\boldsymbol{w} : \|\boldsymbol{w}\| < \sqrt{\frac{4n-1}{4n}}\right\}$, *which satisfies*
>
> $$\|\boldsymbol{w}_\star - \mathbf{0}\| \;\leq\; \min\left\{\frac{c_c \mu}{\theta}\sqrt{\frac{n \log p}{p}}, \frac{\mu}{16}\right\}. \tag{4.1.5}$$
>
> *In particular, with this choice of* p, *the probability the claim fails to hold is at most* $4np^{-10} + \theta(np)^{-7} + \exp\left(-0.3\theta n p\right) + c_d \exp\left(-c_e p \mu^2 \theta^2 / n^2\right)$. *Here* c_a *to* c_e *are all positive absolute constants.*

Here $\boldsymbol{q}(\mathbf{0}) = \boldsymbol{e}_n$, which exactly recovers the last row of $\boldsymbol{X}_0$. Though the unique local minimizer $\boldsymbol{w}_\star$ may not be $\mathbf{0}$, it is very near to $\mathbf{0}$.[1] Hence the resulting $\boldsymbol{q}(\boldsymbol{w}_\star)$ produces a close approximation to $\boldsymbol{x}_0^n$. Note that $\boldsymbol{q}(\Gamma)$ (strictly) contains all points $\boldsymbol{q} \in \mathbb{S}^{n-1}$ such that $n = \arg\max_{i \in \pm[n]} \boldsymbol{q}^* \boldsymbol{e}_i$. We can characterize the graph of the function $f(\boldsymbol{q}; \boldsymbol{X}_0)$ in the vicinity of other signed basis vector $\pm \boldsymbol{e}_i$ simply by changing the plane $\boldsymbol{e}_n^\perp$ to $\boldsymbol{e}_i^\perp$. Doing this $2n$ times (and multiplying the failure probability in Theorem 4.1 by $2n$), we obtain a characterization of $f(\boldsymbol{q}; \boldsymbol{X}_0)$ over the entirety of $\mathbb{S}^{n-1}$.[2] The result is captured by the next corollary.

> **Corollary 4.2** *Suppose* $\boldsymbol{A}_0 = \boldsymbol{I}$ *and hence* $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0 = \boldsymbol{X}_0$. *There exists a positive absolute constant* C, *such that for any* $\theta \in (0, 1/2)$ *and* $\mu < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}$, *whenever* $p \geq \frac{C}{\mu^2\theta^2} n^3 \log \frac{n}{\mu\theta}$, *with probability at least* $1 - 8n^2 p^{-10} - \theta(np)^{-7} - \exp\left(-0.3\theta n p\right) - c_c \exp\left(-c_d p \mu^2 \theta^2 / n^2\right)$, *the function* $f(\boldsymbol{q}; \boldsymbol{X}_0)$ *has exactly* $2n$

[1] As can be seen from the proof, the reason $\boldsymbol{w}_\star$ may not be $\mathbf{0}$ is exactly that p is finite.

[2] In fact, it is possible to pull the very detailed geometry captured in (4.1.2) through (4.1.4) back to the sphere (i.e., the $\boldsymbol{q}$ space) also; analysis of the Riemannian trust-region algorithm later does part of these. We will stick to this simple global version here.

local minimizers over the sphere $\mathbb{S}^{n-1}$. In particular, there is a bijective map between these minimizers and signed basis vectors $\{\pm \boldsymbol{e}_i\}_i$, such that the corresponding local minimizer $\boldsymbol{q}_\star$ and $\boldsymbol{b} \in \{\pm \boldsymbol{e}_i\}_i$ satisfy

$$\|\boldsymbol{q}_\star - \boldsymbol{b}\| \le \sqrt{2}\min\left\{\frac{c_c\mu}{\theta}\sqrt{\frac{n\log p}{p}}, \frac{\mu}{16}\right\}. \tag{4.1.6}$$

Here c_a to c_d are positive absolute constants (possibly different from that in the above theorem).

Proof By Theorem 4.1, over $\boldsymbol{q}(\Gamma)$, $\boldsymbol{q}(\boldsymbol{w}_\star)$ is the unique local minimizer. Suppose not. Then there exist $\boldsymbol{q}' \in \boldsymbol{q}(\Gamma)$ with $\boldsymbol{q}' \neq \boldsymbol{q}(\boldsymbol{w}_\star)$ and $\varepsilon > 0$, such that $f(\boldsymbol{q}'; \boldsymbol{X}_0) \le f(\boldsymbol{q}; \boldsymbol{X}_0)$ for all $\boldsymbol{q} \in \boldsymbol{q}(\Gamma)$ satisfying $\|\boldsymbol{q}' - \boldsymbol{q}\| < \varepsilon$. Since the mapping $\boldsymbol{w} \mapsto \boldsymbol{q}(\boldsymbol{w})$ is $2\sqrt{n}$-Lipschitz (Lemma 9.7), $g(\boldsymbol{w}(\boldsymbol{q}'); \boldsymbol{X}_0) \le g(\boldsymbol{w}(\boldsymbol{q}); \boldsymbol{X}_0)$ for all $\boldsymbol{w} \in \Gamma$ satisfying $\|\boldsymbol{w}(\boldsymbol{q}') - \boldsymbol{w}(\boldsymbol{q})\| < \varepsilon/(2\sqrt{n})$, implying $\boldsymbol{w}(\boldsymbol{q}')$ is a local minimizer different from $\boldsymbol{w}_\star$, a contradiction. Let $\|\boldsymbol{w}_\star - \boldsymbol{0}\| = \eta$. Straightforward calculation shows

$$\|\boldsymbol{q}(\boldsymbol{w}_\star) - \boldsymbol{e}_n\|^2 = \left(1 - \sqrt{1-\eta^2}\right)^2 + \eta^2 = 2 - 2\sqrt{1-\eta^2} \le 2\eta^2.$$

Repeating the argument $2n$ times in the vicinity of other signed basis vectors $\pm\boldsymbol{e}_i$ gives $2n$ local minimizers of f. Indeed, the $2n$ symmetric sections cover the sphere with certain overlaps, and a simple calculation shows that no such local minimizer lies in the overlapped regions (due to nearness to a signed basis vector). There is no extra local minimizer, as such local minimizer is contained in at least one of the $2n$ symmetric sections, resulting two different local minimizers in one section, contradicting the uniqueness result we obtained above. ■

Though the $2n$ isolated local minimizers may have different objective values, they are equally good in the sense any of them produces a close approximation to a certain row of $\boldsymbol{X}_0$. As discussed in Section 3.3.2, for cases $\boldsymbol{A}_0$ is an orthobasis other than $\boldsymbol{I}$, the landscape of $f(\boldsymbol{q}; \boldsymbol{Y})$ is simply a rotated version of the one we characterized above.

The function landscape for general complete $\boldsymbol{A}_0$ is characterized as below.

Theorem 4.3 (High-dimensional landscape - complete dictionary) *Suppose $\boldsymbol{A}_0$ is complete with its condition number $\kappa(\boldsymbol{A}_0)$. There exist positive absolute constants $c_\star$ and C, such that for any $\theta \in (0, 1/2)$ and $\mu < \min\left\{c_a\theta n^{-1}, c_b n^{-5/4}\right\}$, when*

$$p \ge \frac{C}{c_\star^2\theta}\max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\}\kappa^8(\boldsymbol{A}_0)\log^4\left(\frac{\kappa(\boldsymbol{A}_0)\, n}{\mu\theta}\right) \tag{4.1.7}$$

and write $\overline{\boldsymbol{Y}} \doteq \sqrt{p\theta}\left(\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}$, $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathtt{SVD}(\boldsymbol{A}_0)$, the following hold simultaneously with high probabil-

ity:

$$\nabla^2 g(\boldsymbol{w}; \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}}) \succeq \frac{c_\star \theta}{2\mu}\boldsymbol{I} \qquad \forall\, \boldsymbol{w} \quad s.t. \quad \|\boldsymbol{w}\| \le \frac{\mu}{4\sqrt{2}}, \tag{4.1.8}$$

$$\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w}; \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}})}{\|\boldsymbol{w}\|} \ge \frac{1}{2}c_\star\theta \qquad \forall\, \boldsymbol{w} \quad s.t. \quad \frac{\mu}{4\sqrt{2}} \le \|\boldsymbol{w}\| \le \frac{1}{20\sqrt{5}} \tag{4.1.9}$$

$$\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w}; \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}})\boldsymbol{w}}{\|\boldsymbol{w}\|^2} \le -\frac{1}{2}c_\star\theta \qquad \forall\, \boldsymbol{w} \quad s.t. \quad \frac{1}{20\sqrt{5}} \le \|\boldsymbol{w}\| \le \sqrt{\frac{4n-1}{4n}}, \tag{4.1.10}$$

and *the function* $g(\boldsymbol{w}; \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}})$ *has exactly one local minimizer* $\boldsymbol{w}_\star$ *over the open set* $\Gamma \doteq \left\{\boldsymbol{w} : \|\boldsymbol{w}\| < \sqrt{\frac{4n-1}{4n}}\right\}$*, which satisfies*

$$\|\boldsymbol{w}_\star - \boldsymbol{0}\| \;\le\; \frac{\mu}{7}. \tag{4.1.11}$$

In particular, with this choice of p*, the probability the claim fails to hold is at most* $4np^{-10} + \theta(np)^{-7} + \exp(-0.3\theta np) + p^{-8} + c_d \exp\left(-c_e p\mu^2\theta^2/n^2\right)$*. Here* c_a *to* c_e *are all positive absolute constants.*

Corollary 4.4 *Suppose* $\boldsymbol{A}_0$ *is complete with its condition number* $\kappa(\boldsymbol{A}_0)$*. There exist positive absolute constants* $c_\star$ *and* C*, such that for any* $\theta \in (0, 1/2)$ *and* $\mu < \min\left\{c_a\theta n^{-1}, c_b n^{-5/4}\right\}$*, when* $p \ge \frac{C}{c_\star^2\theta}\max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\}\kappa^8(\boldsymbol{A}_0)\log^4\left(\frac{\kappa(\boldsymbol{A}_0)n}{\mu\theta}\right)$ *and* $\overline{\boldsymbol{Y}} \doteq \sqrt{p\theta}\,(\boldsymbol{Y}\boldsymbol{Y}^*)^{-1/2}\boldsymbol{Y}$*,* $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathtt{SVD}(\boldsymbol{A}_0)$*, with probability at least* $1 - 8n^2p^{-10} - \theta(np)^{-7} - \exp(-0.3\theta np) - p^{-8} - c_d\exp\left(-c_e p\mu^2\theta^2/n^2\right)$*, the function* $f\left(\boldsymbol{q}; \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}}\right)$ *has exactly* $2n$ *local minimizers over the sphere* $\mathbb{S}^{n-1}$*. In particular, there is a bijective map between these minimizers and signed basis vectors* $\{\pm\boldsymbol{e}_i\}_i$*, such that the corresponding local minimizer* $\boldsymbol{q}_\star$ *and* $\boldsymbol{b} \in \{\pm\boldsymbol{e}_i\}_i$ *satisfy*

$$\|\boldsymbol{q}_\star - \boldsymbol{b}\| \le \frac{\sqrt{2}\mu}{7}. \tag{4.1.12}$$

Here c_a *to* c_d *are positive absolute constants (possibly different from that in the above theorem).*

We will omit the proof as it is almost identical to that of Corollary 4.2.

From the above results, it is clear that all local minimizers of $f(\boldsymbol{q})$ over $\mathbb{S}^{n-1}$ are "global" in the sense that any of them produces a close approximation to a row of $\boldsymbol{X}_0$ and finding them all approximately recovers all rows of $\boldsymbol{X}_0$. Moreover, any $g(\boldsymbol{w})$ for one of the $2n$ symmetric sections is ridable-saddle with concrete parameters. A natural question is whether $f(\boldsymbol{q})$ is a also ridable-saddle function, which together with the above globalness implies f is *qualitatively* an $\mathcal{X}$ function over the sphere.

The answer is indeed yes. Instead of presenting a rigorous technical statement and detailed proof, we include here just an informal argument. Our analysis of the trust-region algorithm runs back and forth in the $\boldsymbol{w}$ and $\boldsymbol{q}$ space, and it turns out such a lack will not affect our arguments there.

Again we work with one of the symmetric sections first. It is easy to verify the following fact (see proof of Lemma 5.11 on Page 119):

$$\langle \operatorname{grad} f(\boldsymbol{q}), \boldsymbol{q} - \boldsymbol{e}_n / q_n \rangle = \langle \boldsymbol{w}, \nabla g(\boldsymbol{w}) \rangle .$$

Thus, $\langle \operatorname{grad} f(\boldsymbol{q}), \boldsymbol{q} - \boldsymbol{e}_n / q_n \rangle \neq 0$ if and only if $\langle \boldsymbol{w}, \nabla g(\boldsymbol{w}) \rangle \neq 0$, implying that $\operatorname{grad} f(\boldsymbol{q})$ will never be zero in $\left\{ \boldsymbol{q}(\boldsymbol{w}) : \mu/(4\sqrt{2}) \le \|\boldsymbol{w}\| \le 1/(20\sqrt{5}) \right\}$. Moreover, it is shown in Lemma 5.9 below that the Riemannian Hessian is positive definite for any point in the spherical region $\left\{ \boldsymbol{q}(\boldsymbol{w}) : \|\boldsymbol{w}\| \le \mu/(4\sqrt{2}) \right\}$. Moreover, Theorem 4.1 and Theorem 4.3 imply that around each point in $\{\boldsymbol{w} : 1/(20\sqrt{5}) \le \|\boldsymbol{w}\| \le \sqrt{(4n-1)/(4n)}\}$, $g(\boldsymbol{w})$ is *strictly* concave locally in $\pm\boldsymbol{w}$ direction. Intuitively, through the $\boldsymbol{q}(\boldsymbol{w})$ mapping, the function $\boldsymbol{f}(\boldsymbol{q})$ is geodesically strictly concave along the geodesic curve connecting $\pm\boldsymbol{e}_n$. Repeating the above arguments for all $2n$ symmetric sections, and then noting any point on the sphere is covered by one out of the strong gradient, strongly convex, and negative directional curvature regions, we can conclude that $f(\boldsymbol{q})$ is indeed a ridable-saddle function (with concrete parameters that we have not estimated), and hence also an "approximate" $\mathcal{X}$ function over the sphere.

4.2 Sketch of proof ideas for orthogonal dictionaries

The proof of Theorem 4.1 is conceptually straightforward: one shows that $\mathbb{E}_{\boldsymbol{X}_0}\left[g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)\right]$ has the claimed properties, and then proves that each of the quantities of interest concentrates uniformly about its expectation. The detailed calculations are nontrivial.

The next three propositions show that in the expected function landscape, we see successively strongly convex region, nonzero gradient region, and directional negative curvature region when moving away from zero, as depicted in Figure 3.2 and sketched in Section 3.3.2. Note that in this case

$$\mathbb{E}_{\boldsymbol{X}_0}\left[g\left(\boldsymbol{q}; \boldsymbol{X}_0\right)\right] = \mathbb{E}_{\boldsymbol{x} \sim_{i.i.d.} \mathrm{BG}(\theta)}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right].$$

Proposition 4.5 *There exists an absolute constant $c > 0$, such that for every $\theta \in (0, 1/2)$ and any $R_h \in \left(0, \sqrt{\frac{4n-1}{4n}}\right)$, if $\mu \le c \min\left\{\theta R_h^2 n^{-1}, R_h n^{-5/4}\right\}$, it holds for every $\boldsymbol{w}$ satisfying $R_h \le \|\boldsymbol{w}\| \le \sqrt{\frac{4n-1}{4n}}$ that*

$$\frac{\boldsymbol{w}^* \nabla^2_{\boldsymbol{w}} \mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right]\boldsymbol{w}}{\|\boldsymbol{w}\|^2} \le -\frac{\theta}{2\sqrt{2\pi}}.$$

Proof See Section 9.1.1 on Page 80. ■

> **Proposition 4.6** *For every $\theta \in (0, 1/2)$ and every $\mu \le 9/50$, it holds for every $\boldsymbol{w}$ satisfying $r_g \le \|\boldsymbol{w}\| \le R_g$, where $r_g = \mu/(6\sqrt{2})$ and $R_g = (1-\theta)/(10\sqrt{5})$, that*
>
> $$\frac{\boldsymbol{w}^* \nabla_{\boldsymbol{w}} \mathbb{E}\left[h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x})\right]}{\|\boldsymbol{w}\|} \ge \frac{\theta}{20\sqrt{2\pi}}.$$

Proof See Section 9.1.2 on Page 86. ■

> **Proposition 4.7** *For every $\theta \in (0, 1/2)$, and every $\mu \le 1/(20\sqrt{n})$, it holds for every $\boldsymbol{w}$ satisfying $\|\boldsymbol{w}\| \le \mu/(4\sqrt{2})$ that*
>
> $$\mathbb{E}\left[\nabla^2_{\boldsymbol{w}} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] \succeq \frac{\theta}{25\sqrt{2\pi}\mu}\boldsymbol{I}.$$

Proof See Section 9.1.3 on Page 88. ■

To prove that the above hold qualitatively for finite p, i.e., the function $g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)$, we will first prove that for a fixed $\boldsymbol{w}$ each of the quantity of interest concentrate about their expectation w.h.p., and the function is nice enough (Lipschitz) such that we can extend the results to all $\boldsymbol{w}$ via a discretization argument. The next three propositions provide the desired pointwise concentration results.

> **Proposition 4.8** *Suppose $0 < \mu \le 1/\sqrt{n}$. For every $\boldsymbol{w} \in \Gamma$, it holds that for any $t > 0$,*
>
> $$\mathbb{P}\left[\left|\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0)\boldsymbol{w}}{\|\boldsymbol{w}\|^2} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0)\boldsymbol{w}}{\|\boldsymbol{w}\|^2}\right]\right| \ge t\right] \le 4\exp\left(-\frac{p\mu^2 t^2}{512 n^2 + 32 n\mu t}\right).$$

Proof See Page 92 under Section 9.1.4. ■

> **Proposition 4.9** *For every $\boldsymbol{w} \in \Gamma$, it holds that for any $t > 0$,*
>
> $$\mathbb{P}\left[\left|\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w}; \boldsymbol{X}_0)}{\|\boldsymbol{w}\|} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w}; \boldsymbol{X}_0)}{\|\boldsymbol{w}\|}\right]\right| \ge t\right] \le 2\exp\left(-\frac{p t^2}{8n + 4\sqrt{n} t}\right).$$

Proof See Page 93 under Section 9.1.4. ■

> **Proposition 4.10** *Suppose $0 < \mu \le 1/\sqrt{n}$. For every $\boldsymbol{w} \in \Gamma \cap \{\boldsymbol{w} : \|\boldsymbol{w}\| \le 1/4\}$, it holds that for any $t > 0$,*
>
> $$\mathbb{P}\left[\left\|\nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0) - \mathbb{E}\left[\nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0)\right]\right\| \ge t\right] \ \le\ 4n\exp\left(-\frac{p\mu^2 t^2}{512 n^2 + 32\mu n t}\right).$$

Proof See Page 93 under Section 9.1.4. ■

The next three propositions provide the desired Lipschitz results. Here $\|\cdot\|_\infty$ returns the largest of elementwise magnitudes.

Proposition 4.11 (Hessian Lipschitz) *Fix any* $r_{\frown} \in (0,1)$*. Over the set* $\Gamma \cap \{\boldsymbol{w} : \|\boldsymbol{w}\| \geq r_{\frown}\}$, $\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0) \boldsymbol{w}}{\|\boldsymbol{w}\|^2}$ *is* $L_{\frown}$*-Lipschitz with*

$$L_{\frown} \leq \frac{16n^3}{\mu^2} \|\boldsymbol{X}_0\|_\infty^3 + \frac{8n^{3/2}}{\mu r_{\frown}} \|\boldsymbol{X}_0\|_\infty^2 + \frac{48 n^{5/2}}{\mu} \|\boldsymbol{X}_0\|_\infty^2 + 96 n^{5/2} \|\boldsymbol{X}_0\|_\infty .$$

Proof See Page 99 under Section 9.1.5. ■

Proposition 4.12 (Gradient Lipschitz) *Fix any* $r_g \in (0,1)$*. Over the set* $\Gamma \cap \{\boldsymbol{w} : \|\boldsymbol{w}\| \geq r_g\}$, $\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w}; \boldsymbol{X}_0)}{\|\boldsymbol{w}\|}$ *is* L_g*-Lipschitz with*

$$L_g \leq \frac{2\sqrt{n} \|\boldsymbol{X}_0\|_\infty}{r_g} + 8n^{3/2} \|\boldsymbol{X}_0\|_\infty + \frac{4n^2}{\mu} \|\boldsymbol{X}_0\|_\infty^2 .$$

Proof See Page 99 under Section 9.1.5. ■

Proposition 4.13 (Lipschitz for Hessian around zero) *Fix any* $r_{\smile} \in \left(0, \frac{1}{2}\right)$*. Over the set* $\Gamma \cap \{\boldsymbol{w} : \|\boldsymbol{w}\| \leq r_{\smile}\}$, $\nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0)$ *is* $L_{\smile}$*-Lipschitz with*

$$L_{\smile} \;\leq\; \frac{4n^2}{\mu^2} \|\boldsymbol{X}_0\|_\infty^3 + \frac{4n}{\mu} \|\boldsymbol{X}_0\|_\infty^2 + \frac{8\sqrt{2}\sqrt{n}}{\mu} \|\boldsymbol{X}_0\|_\infty^2 + 8 \|\boldsymbol{X}_0\|_\infty .$$

Proof See Page 100 under Section 9.1.5. ■

Integrating the above pieces, Section 9.2 provides a complete proof of Theorem 4.1.

4.3 Extending to complete dictionaries

As hinted in Section 3.3.2, instead of proving things from scratch, we build on the results we have obtained for orthogonal dictionaries. In particular, we will work with the preconditioned data matrix

$$\overline{\boldsymbol{Y}} \doteq \sqrt{p\theta} \left(\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2} \boldsymbol{Y} \tag{4.3.1}$$

and show that the function landscape $f\left(\boldsymbol{q}; \overline{\boldsymbol{Y}}\right)$ looks qualitatively like that of orthogonal dictionaries (up to a global rotation), provided that p is large enough.

The next lemma shows $\overline{\boldsymbol{Y}}$ can be treated as being generated from an orthobasis with the same BG coefficients, plus small noise.

> **Lemma 4.14** *For any $\theta \in (0, 1/2)$, suppose $\boldsymbol{A}_0$ is complete with condition number $\kappa(\boldsymbol{A}_0)$ and $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}(\theta)$. Provided $p \geq C\kappa^4(\boldsymbol{A}_0)\,\theta n^2 \log(n\theta\kappa(\boldsymbol{A}_0))$, one can write $\overline{\boldsymbol{Y}}$ as defined in (4.3.1) as*
>
> $$\overline{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \boldsymbol{\Xi}\boldsymbol{X}_0,$$
>
> *for a certain $\boldsymbol{\Xi}$ obeying $\|\boldsymbol{\Xi}\| \leq 20\kappa^4(\boldsymbol{A})\sqrt{\frac{\theta n \log p}{p}}$, with probability at least $1 - p^{-8}$. Here $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathtt{SVD}(\boldsymbol{A}_0)$, and $C > 0$ is an absolute constant.*

Proof See Page 104 under Section 9.3. ■

Notice that $\boldsymbol{U}\boldsymbol{V}^*$ above is orthogonal, and that landscape of $f(\boldsymbol{q};\overline{Y})$ is simply a rotated version of that of $f(\boldsymbol{q};\boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}})$, or using the notation in the above lemma, that of $f(\boldsymbol{q};\boldsymbol{X}_0 + \boldsymbol{V}\boldsymbol{U}^*\boldsymbol{\Xi}\boldsymbol{X}_0) = f(\boldsymbol{q};\boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0)$ assuming $\widetilde{\boldsymbol{\Xi}} \doteq \boldsymbol{V}\boldsymbol{U}^*\boldsymbol{\Xi}$. So similar to the orthogonal case, it is enough to consider this "canonical" case, and its "canonical" reparametrization:

$$g\left(\boldsymbol{w};\boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) = \frac{1}{p}\sum_{k=1}^{p} h_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})(\boldsymbol{x}_0)_k + \boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{\Xi}}(\boldsymbol{x}_0)_k\right).$$

The following lemma provides quantitative comparison between the gradient and Hessian of $g\left(\boldsymbol{w};\boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right)$ and that of $g(\boldsymbol{w};\boldsymbol{X}_0)$.

> **Lemma 4.15** *There exist positive constants C_a and C_b, such that for all $\boldsymbol{w} \in \Gamma$,*
>
> $$\left\|\nabla_{\boldsymbol{w}} g(\boldsymbol{w};\boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0) - \nabla_{\boldsymbol{w}} g(\boldsymbol{w};\boldsymbol{X}_0)\right\| \leq C_a \frac{n}{\mu}\log(np)\,\|\widetilde{\boldsymbol{\Xi}}\|,$$
> $$\left\|\nabla^2_{\boldsymbol{w}} g(\boldsymbol{w};\boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0) - \nabla^2_{\boldsymbol{w}} g(\boldsymbol{w};\boldsymbol{X}_0)\right\| \leq C_b \max\left\{\frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu}\right\}\log^{3/2}(np)\,\|\widetilde{\boldsymbol{\Xi}}\|$$
>
> *with probability at least $1 - \theta(np)^{-7} - \exp(-0.3\theta np)$.*

Proof See Page 105 under Section 9.3. ■

Combining the above two lemmas, it is easy to see when p is large enough, $\|\widetilde{\boldsymbol{\Xi}}\| = \|\boldsymbol{\Xi}\|$ is then small enough (Lemma 4.14), and hence the changes to the gradient and Hessian caused by the perturbation are small. This gives the results presented in Theorem 4.3; see Section 9.3 for a detailed proof. In particular, for the p chosen in Theorem 4.3, it holds that

$$\left\|\widetilde{\boldsymbol{\Xi}}\right\| \leq cc_\star\theta\left(\max\left\{\frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu}\right\}\log^{3/2}(np)\right)^{-1} \tag{4.3.2}$$

for a constant $c > 0$ which can be made arbitrarily small by making the constant C in p large.

Chapter 5

Finding One Local Minimizer via the Riemannian Trust-Region Method

> Nevertheless, it remains conceivable that the measure relations of space in the infinitely small are not in accordance with the assumptions of our geometry [Euclidean geometry], and, in fact, we should have to assume that they are not if, by doing so, we should ever be enabled to explain phenomena in a more simple way.
>
> Bernhard Riemann

The geometric results in the preceding chapter show that each local minimizer of $f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})$ over $\mathbb{S}^{n-1}$ approximately recovers a row of $\boldsymbol{X}_0$ and by finding all local minimizers one can recover all rows of $\boldsymbol{X}_0$. So the central question left is how to efficiently obtain these local minimizers. In this chapter, we focus on finding any one local minimizer out of the many; we will discuss how to sequentially find more based on the result here and hence to recover $\boldsymbol{X}_0$ in the next chapter. The presence of saddle points has motivated us to develop a (second-order) Riemannian trust-region algorithm over the sphere; the existence of descent directions at nonoptimal points drives the trust-region iteration sequence towards one of the minimizers asymptotically. We will prove that under our model assumptions, the algorithm efficiently produces a close approximation (up to numerical precision) to one of the minimizers. Throughout the exposition, basic knowledge of Riemannian geometry is assumed. We will try to keep the technical requirement minimal possible; the reader can consult the excellent monograph [AMS09] for relevant background and details.

We will provide a self-contained development of trust-region method over the sphere in Section 5.1 with implementation details, followed by main convergence results in Section 5.2. We will then sketch how to prove the convergence results for orthogonal and complete dictionaries in Section 5.3 and Section 5.4, respectively. Details proofs are deferred to Chapter 10.

5.1 The Riemannian trust-region algorithm over the sphere

We are interested to seek one local minimizer of the problem

$$\mathrm{minimize} \quad f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \doteq \frac{1}{p}\sum_{k=1}^{p} h_\mu(\boldsymbol{q}^* \widehat{\boldsymbol{y}}_i) \quad \text{subject to} \quad \boldsymbol{q} \in \mathbb{S}^{n-1}. \tag{5.1.1}$$

For a function $f(\boldsymbol{q})$ in the Euclidean space, the typical TRM starts from some initialization $\boldsymbol{q}^{(0)} \in \mathbb{R}^n$, and produces a sequence of iterates $\boldsymbol{q}^{(1)}, \boldsymbol{q}^{(2)}, \dots$, by repeatedly minimizing a quadratic approximation $\widehat{f}$ to the objective function $f(\boldsymbol{q})$, over a ball centered about the current iterate.

Here, we are interested in the restriction of $f(\boldsymbol{q})$ to the unit sphere $\mathbb{S}^{n-1}$. Instead of directly approximating the function in $\mathbb{R}^n$, we form quadratic approximations of f in the tangent space of $\mathbb{S}^{n-1}$. Recall that the tangent space of a sphere at a point $\boldsymbol{q} \in \mathbb{S}^{n-1}$ is $T_{\boldsymbol{q}}\mathbb{S}^{n-1} = \{\boldsymbol{\delta} \in \mathbb{R}^n : \boldsymbol{q}^*\boldsymbol{\delta} = 0\}$, i.e., the set of vectors that are orthogonal to $\boldsymbol{q}$. Consider the exponential map $\exp_{\boldsymbol{q}}(\delta) \doteq \boldsymbol{q} \cos \|\boldsymbol{\delta}\| + \boldsymbol{\delta}/\|\boldsymbol{\delta}\| \cdot \sin \|\boldsymbol{\delta}\|$ that maps a neighboring point $\boldsymbol{\delta}$ of $\mathbf{0}$ on $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ to a point near $\boldsymbol{q}$ on $\mathbb{S}^{n-1}$. The function $f \circ \exp_{\boldsymbol{q}}(\boldsymbol{\delta})$ obviously is smooth and we expect Taylor expansion around 0 a good approximation of the function, at least in the vicinity of $\mathbf{0}$. Taylor's theorem gives

$$f \circ \exp_{\boldsymbol{q}}(\boldsymbol{\delta}) = f(\boldsymbol{q}) + \left\langle \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \boldsymbol{\delta} \right\rangle + \frac{1}{2}\boldsymbol{\delta}^* \left(\nabla^2 f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \boldsymbol{q} \right\rangle \boldsymbol{I} \right) \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3).$$

We therefore form the "quadratic" approximation $\widehat{f}(\boldsymbol{\delta}; \boldsymbol{q}) : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \mapsto \mathbb{R}$ as

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{q}, \widehat{\boldsymbol{Y}}) \;\doteq\; f(\boldsymbol{q}) + \left\langle \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \boldsymbol{\delta} \right\rangle + \frac{1}{2}\boldsymbol{\delta}^* \left(\nabla^2 f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \boldsymbol{q} \right\rangle \boldsymbol{I} \right) \boldsymbol{\delta}. \tag{5.1.2}$$

Given the previous iterate $\boldsymbol{q}^{(k-1)}$, the TRM produces the next iterate by generating a solution $\widehat{\boldsymbol{\delta}}$ to

$$\mathrm{minimize}_{\boldsymbol{\delta} \in T_{\boldsymbol{q}^{(k-1)}}\mathbb{S}^{n-1},\, \|\boldsymbol{\delta}\| \le \Delta} \quad \widehat{f}(\boldsymbol{\delta}; \boldsymbol{q}^{(k-1)}), \tag{5.1.3}$$

and then "pull" the solution $\widehat{\boldsymbol{\delta}}$ from $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ back to $\mathbb{S}^{n-1}$. If we choose the exponential map to pull back the movement $\widehat{\boldsymbol{\delta}}$,[1] the next iterate then reads

$$\boldsymbol{q}^{(k)} = \boldsymbol{q}^{(k-1)} \cos \|\widehat{\boldsymbol{\delta}}\| + \frac{\widehat{\boldsymbol{\delta}}}{\|\widehat{\boldsymbol{\delta}}\|} \sin \|\widehat{\boldsymbol{\delta}}\|. \tag{5.1.4}$$

We have motivated (5.1.2) and hence the algorithm in an intuitive way from the Taylor approximation to the function f over $\mathbb{S}^{n-1}$. To understand its properties, it is useful to interpret it as a *Riemannian trust-region method* over the manifold $\mathbb{S}^{n-1}$. The class of algorithms are discussed in detail in the monograph [AMS09]. In particular, the quadratic approximation (5.1.2) can be obtained by noting that the function $f \circ \exp_{\boldsymbol{q}}(\boldsymbol{\delta}; \widehat{\boldsymbol{Y}}) : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \mapsto \mathbb{R}$ obeys

$$f \circ \exp_{\boldsymbol{q}}(\boldsymbol{\delta}; \widehat{\boldsymbol{Y}}) = f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) + \left\langle \boldsymbol{\delta}, \operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \right\rangle + \frac{1}{2}\boldsymbol{\delta}^* \operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})\boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3),$$

where $\operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})$ and $\operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})$ are the Riemannian gradient and Riemannian Hessian [AMS09] respectively, defined as

$$\begin{aligned} \operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) &\doteq \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}} \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \\ \operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) &\doteq \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}} \left(\nabla^2 f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \boldsymbol{q} \right\rangle \boldsymbol{I} \right) \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}, \end{aligned}$$

with $\mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}} \doteq \boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^*$ the orthoprojector onto the tangent space $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$. We will use these standard notions in analysis of the algorithm.

To solve the subproblem (5.1.3) numerically, we can take any matrix $\boldsymbol{U} \in \mathbb{R}^{n \times (n-1)}$ whose columns form an orthonormal basis for $T_{\boldsymbol{q}^{(k-1)}}\mathbb{S}^{n-1}$, and produce a solution $\widehat{\boldsymbol{\xi}}$ to

$$\operatorname{minimize}_{\|\boldsymbol{\xi}\| \le \Delta} \quad \widehat{f}(\boldsymbol{U}\boldsymbol{\xi}; \boldsymbol{q}^{(k-1)}), \tag{5.1.5}$$

where by (5.1.2),

$$\begin{aligned} \widehat{f}(\boldsymbol{U}\boldsymbol{\xi}; \boldsymbol{q}^{(k-1)}) = f(\boldsymbol{q}) + \left\langle \boldsymbol{U}^* \nabla f(\boldsymbol{q}^{(k-1)}), \boldsymbol{\xi} \right\rangle + \\ \frac{1}{2}\boldsymbol{\xi}^* \left(\boldsymbol{U}^* \nabla^2 f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}})\boldsymbol{U} - \left\langle \nabla f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}}), \boldsymbol{q}^{(k-1)} \right\rangle \boldsymbol{I}_{n-1} \right) \boldsymbol{\xi}. \end{aligned}$$

Solution to (5.1.3) can then be recovered as $\widehat{\boldsymbol{\delta}} = \boldsymbol{U}\widehat{\boldsymbol{\xi}}$. The problem (5.1.5) is an instance of the classic *trust region subproblem*, i.e., minimizing a quadratic function over an ℓ^2 norm ball, which can be solved in polynomial time, either by root finding methods [MS83, CGT00] or by semidefinite programming (SDP) [RW97, YZ03, FW04, HK14]. We brief discuss how SDP can be used to solve the subproblem exactly here; our implementation in

[1] The exponential map is only one of the many possibilities; also for general manifolds other retraction schemes may be more practical. See exposition on retraction in Chapter 4 of [AMS09].

simulations are based on the modified Manopt implementation discussed in Section 2.6. We introduce

$$\tilde{\boldsymbol{\xi}} = [\boldsymbol{\xi}^*, 1]^*, \; \boldsymbol{\Theta} = \tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}^*, \; \boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^* & 0 \end{bmatrix}, \tag{5.1.6}$$

where $\boldsymbol{A} = \boldsymbol{U}^*(\nabla^2 f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}}), \boldsymbol{q}^{(k-1)} \right\rangle \boldsymbol{I})\boldsymbol{U}$ and $\boldsymbol{b} = \boldsymbol{U}^* \nabla f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}})$. The resulting SDP to solve is

$$\text{minimize}_{\,\boldsymbol{\Theta}} \, \langle \boldsymbol{M}, \boldsymbol{\Theta} \rangle, \; \text{subject to} \; \operatorname{tr}(\boldsymbol{\Theta}) \le \Delta^2 + 1, \; \langle \boldsymbol{E}_{n+1}, \boldsymbol{\Theta} \rangle = 1, \; \boldsymbol{\Theta} \succeq \boldsymbol{0}, \tag{5.1.7}$$

where $\boldsymbol{E}_{n+1} = \boldsymbol{e}_{n+1}\boldsymbol{e}_{n+1}^*$. Once the problem (5.1.7) is solved to its optimal $\boldsymbol{\Theta}_\star$, one can provably recover the optimal solution $\boldsymbol{\xi}_\star$ of (5.1.5) by computing the SVD of $\boldsymbol{\Theta}_\star = \widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{V}}^*$, and extract as a subvector by the first $n-1$ coordinates of the principal eigenvector $\widetilde{\boldsymbol{u}}_1$ (see Appendix B of [BV04]).

5.2 Main convergence results

By specializing general results on the Riemannian TRM (see, e.g., Chapter 7 of [AMS09]), it is not difficult to prove that the iterates sequence $\boldsymbol{q}^{(k)}$ described above converges to a critical point of the objective $f(\boldsymbol{q})$ over $\mathbb{S}^{n-1}$. In this section, we show that under our probabilistic assumptions, a stronger result can be obtained (see also [BAC16]). Specifically, the iterative algorithm is guaranteed to produce a close approximation to a local minimizer of the objective function, in a number of iterations that is polynomial in the problem size. The arguments described in Chapter 4 show that with high probability every local minimizer of f produces a close approximation of one row of $\boldsymbol{X}_0$. Taken together, this implies that the algorithm efficiently produces a close approximation to one row of $\boldsymbol{X}_0$.

Our next two theorems summarize the convergence results for orthogonal and complete dictionaries, respectively.

> **Theorem 5.1 (TRM convergence - orthogonal dictionary)** *Suppose the dictionary $\boldsymbol{A}_0$ is orthogonal. There exists a positive constant C, such that for all $\theta \in (0, 1/2)$, and $\mu < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}$, whenever*
>
> $$\exp(n) \ge p \ge C n^3 \log \tfrac{n}{\mu\theta} / (\mu^2\theta^2), \tag{5.2.1}$$
>
> *with probability at least $1 - 8n^2 p^{-10} - \theta(np)^{-7} - \exp(-0.3\theta n p) - p^{-10} - c_c \exp\left(-c_d p \mu^2 \theta^2 / n^2\right)$, the Riemannian trust-region algorithm with input data matrix $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}$, any initialization $\boldsymbol{q}^{(0)}$ on the sphere, and a step*

size satisfying

$$\Delta \le \min \left\{ \frac{c_e c_\star \theta \mu^2}{n^{5/2} \log^{3/2}(np)}, \frac{c_f c_\sharp^3 \theta^3 \mu}{n^{7/2} \log^{7/2}(np)} \right\}, \tag{5.2.2}$$

returns a solution $\widehat{\boldsymbol{q}} \in \mathbb{S}^{n-1}$ *which is* ε *near to one of the local minimizers* $\boldsymbol{q}_\star$ *(i.e.,* $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \le \varepsilon$*) in*

$$\max \left\{ \frac{c_g n^6 \log^3(np)}{c_\star^3 \theta^3 \mu^4}, \frac{c_h n}{c_\sharp^2 \theta^2 \Delta^2} \right\} f(\boldsymbol{q}^{(0)}) + \log\log \frac{c_i c_\star \theta \mu}{\varepsilon n^{3/2} \log^{3/2}(np)} \tag{5.2.3}$$

iterations. Here $c_\star$*,* $c_\sharp$ *as defined in Theorem 4.1 and Lemma 5.9 respectively (*$c_\star$ *and* $c_\sharp$ *can be set to the same constant value), and* c_a*,* c_b *are the same constants as defined in Theorem 4.1,* c_c *through* c_i *are other positive constants.*

Theorem 5.2 (TRM convergence - complete dictionary) *Suppose the dictionary* $\boldsymbol{A}_0$ *is complete with condition number* $\kappa(\boldsymbol{A}_0)$*. There exists a positive constant* C*, such that for all* $\theta \in (0, 1/2)$*, and* $\mu < \min\{c_a \theta n^{-1}, c_b n^{-5/4}\}$*, whenever*

$$\exp(n) \ge p \ge \frac{C}{c_\star^2 \theta} \max\left\{ \frac{n^4}{\mu^4}, \frac{n^5}{\mu^2} \right\} \kappa^8(\boldsymbol{A}_0) \log^4\left(\frac{\kappa(\boldsymbol{A}_0) n}{\mu\theta} \right), \tag{5.2.4}$$

with probability at least $1 - 8n^2 p^{-10} - \theta(np)^{-7} - \exp(-0.3\theta np) - 2p^{-8} - c_c \exp(-c_d p \mu^2 \theta^2 / n^2)$*, the Riemannian trust-region algorithm with input data matrix* $\overline{\boldsymbol{Y}} \doteq \sqrt{p\theta}\,(\boldsymbol{Y}\boldsymbol{Y}^*)^{-1/2}\boldsymbol{Y}$ *where* $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathtt{SVD}(\boldsymbol{A}_0)$*, any initialization* $\boldsymbol{q}^{(0)}$ *on the sphere and a step size satisfying*

$$\Delta \le \min \left\{ \frac{c_e c_\star \theta \mu^2}{n^{5/2} \log^{3/2}(np)}, \frac{c_f c_\sharp^3 \theta^3 \mu}{n^{7/2} \log^{7/2}(np)} \right\} \tag{5.2.5}$$

returns a solution $\widehat{\boldsymbol{q}} \in \mathbb{S}^{n-1}$ *which is* ε *near to one of the local minimizers* $\boldsymbol{q}_\star$ *(i.e.,* $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \le \varepsilon$*) in*

$$\max \left\{ \frac{c_g n^6 \log^3(np)}{c_\star^3 \theta^3 \mu^4}, \frac{c_h n}{c_\sharp^2 \theta^2 \Delta^2} \right\} f(\boldsymbol{q}^{(0)}) + \log\log \frac{c_i c_\star \theta \mu}{\varepsilon n^{3/2} \log^{3/2}(np)} \tag{5.2.6}$$

iterations. Here $c_\star$*,* $c_\sharp$ *as defined in Theorem 4.1 and Lemma 5.9 respectively (*$c_\star$ *and* $c_\sharp$ *can be set to the same constant value), and* c_a*,* c_b *are the same constants as defined in Theorem 4.1,* c_c *through* c_i *are other positive constants.*

Our convergence results show that for any target accuracy $\varepsilon > 0$ the algorithm terminates within polynomially many steps. Our estimate of the number of steps is pessimistic: our analysis has assumed a fixed step size Δ and the running time is a relatively large degree polynomial in p and n, while on typical numerical examples the algorithm with adaptive step size produces a solution in relatively few (~ 100) iterations. Nevertheless, our goal in stating the above results is not to provide a tight analysis, but to prove that the Riemannian TRM algorithm finds a local minimizer in polynomial time. For nonconvex problems, this is not entirely trivial – results of [MK87] show that in general it is NP-hard to find a local minimizer of a nonconvex function.

5.3 Useful technical results and sketch of proof for orthogonal dictionaries

The reason that our algorithm is successful derives from the geometry depicted in Figure 3.2 and formalized in Theorem 4.1. Basically, the sphere $\mathbb{S}^{n-1}$ can be divided into three regions. Near each local minimizer, the function is strongly convex, and the algorithm behaves like a standard (Euclidean) TRM algorithm applied to a strongly convex function – in particular, it exhibits a quadratic asymptotic rate of convergence. Away from local minimizers, the function always exhibits either a strong gradient, or a direction of negative curvature (an eigenvalue of the Hessian which is bounded below zero). The Riemannian TRM algorithm is capable of exploiting these quantities to reduce the objective value by at least a fixed amount in each iteration. The total number of iterations spent away from the vicinity of the local minimizers can be bounded by comparing this constant to the initial objective value. Our proofs follow exactly this line and make the various quantities precise.

5.3.1 Basic facts about the sphere

For any point $\boldsymbol{q} \in \mathbb{S}^{n-1}$, the tangent space $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ and the orthoprojector $\mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}$ onto $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ are given by

$$\begin{aligned} T_{\boldsymbol{q}}\mathbb{S}^{n-1} &= \{\boldsymbol{\delta} \in \mathbb{R}^n : \boldsymbol{q}^*\boldsymbol{\delta} = 0\}, \\ \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}} &= \boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^* = \boldsymbol{U}\boldsymbol{U}^*, \end{aligned}$$

where $\boldsymbol{U} \in \mathbb{R}^{n\times(n-1)}$ is an arbitrary orthonormal basis for $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ (note that the orthoprojector is independent of the basis $\boldsymbol{U}$ we choose). Moreover, for any $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$, the exponential map $\exp_{\boldsymbol{q}}(\boldsymbol{\delta}) : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \mapsto \mathbb{S}^{n-1}$ is given by

$$\exp_{\boldsymbol{q}}(\boldsymbol{\delta}) = \boldsymbol{q}\cos\|\boldsymbol{\delta}\| + \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}\sin\|\boldsymbol{\delta}\|.$$

Let $\nabla f(\boldsymbol{q})$ and $\nabla^2 f(\boldsymbol{q})$ denote the usual (Euclidean) gradient and Hessian of f w.r.t. $\boldsymbol{q}$ in $\mathbb{R}^n$. For our specific f defined in (5.1.1), it is easy to check that

$$\nabla f\left(\boldsymbol{q}; \widehat{\boldsymbol{Y}}\right) = \frac{1}{p}\sum_{k=1}^{p} \tanh\left(\frac{\boldsymbol{q}^* \widehat{\boldsymbol{y}}_k}{\mu}\right) \widehat{\boldsymbol{y}}_k, \tag{5.3.1}$$

$$\nabla^2 f\left(\boldsymbol{q}; \widehat{\boldsymbol{Y}}\right) = \frac{1}{p}\sum_{k=1}^{p} \frac{1}{\mu}\left[1 - \tanh^2\left(\frac{\boldsymbol{q}^* \widehat{\boldsymbol{y}}_k}{\mu}\right)\right] \widehat{\boldsymbol{y}}_k \widehat{\boldsymbol{y}}_k^*. \tag{5.3.2}$$

Since $\mathbb{S}^{n-1}$ is an embedded submanifold of $\mathbb{R}^n$, the Riemannian gradient and Riemannian Hessian defined on $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ are given by

$$\operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) = \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}} \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \tag{5.3.3}$$

$$\operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) = \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}} \left(\nabla^2 f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \boldsymbol{q}\right\rangle \boldsymbol{I}\right) \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}; \tag{5.3.4}$$

so the second-order Taylor approximation for the function f is

$$\widehat{f}\left(\boldsymbol{\delta}; \boldsymbol{q}, \widehat{\boldsymbol{Y}}\right) = f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) + \left\langle \boldsymbol{\delta}, \operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})\right\rangle + \frac{1}{2}\boldsymbol{\delta}^* \operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})\boldsymbol{\delta}, \qquad \forall\, \boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}.$$

The first order necessary condition for *unconstrained* minimization of function $\widehat{f}$ over $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ is

$$\operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) + \operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})\boldsymbol{\delta}_\star = \mathbf{0}; \tag{5.3.5}$$

if $\operatorname{Hess} f(\boldsymbol{q})$ is positive semidefinite and has full rank $n-1$ (hence "non-degenerate"[2]), the unique solution $\boldsymbol{\delta}_\star$ is

$$\boldsymbol{\delta}_\star = -\boldsymbol{U}\left(\boldsymbol{U}^* \left[\operatorname{Hess} f(\boldsymbol{q})\right] \boldsymbol{U}\right)^{-1} \boldsymbol{U}^* \operatorname{grad} f(\boldsymbol{q}),$$

which is also invariant to the choice of basis $\boldsymbol{U}$. Given a tangent vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$, let $\gamma(t) \doteq \exp_{\boldsymbol{q}}(t\boldsymbol{\delta})$ denote a geodesic curve on $\mathbb{S}^{n-1}$. Following the notation of [AMS09], let

$$\mathcal{P}_{\gamma}^{\tau \leftarrow 0} : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \to T_{\gamma(\tau)}\mathbb{S}^{n-1}$$

denotes the parallel translation operator, which translates the tangent vector $\boldsymbol{\delta}$ at $\boldsymbol{q} = \gamma(0)$ to a tangent vector at $\gamma(\tau)$, in a "parallel" manner. In the sequel, we identify $\mathcal{P}_{\gamma}^{\tau \leftarrow 0}$ with the following $n \times n$ matrix, whose restriction to $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ is the parallel translation operator (the detailed derivation can be found in Chapter 8.1

[2] Note that the $n \times n$ matrix $\operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})$ has rank at most $n-1$, as the nonzero $\boldsymbol{q}$ obviously is in its null space. When $\operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})$ has rank $n-1$, it has no null direction in the tangent space. Thus, in this case it acts on the tangent space like a full-rank matrix.

of [AMS09]):

$$\begin{aligned} \mathcal{P}_{\gamma}^{\tau \leftarrow 0} &= \left(\boldsymbol{I} - \frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2} \right) - \boldsymbol{q} \sin\left(\tau \|\boldsymbol{\delta}\|\right) \frac{\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|} + \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|} \cos\left(\tau \|\boldsymbol{\delta}\|\right) \frac{\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|} \\ &= \boldsymbol{I} + \left(\cos(\tau \|\boldsymbol{\delta}\|) - 1\right) \frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2} - \sin\left(\tau \|\boldsymbol{\delta}\|\right) \frac{\boldsymbol{q}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|}. \end{aligned} \tag{5.3.6}$$

Similarly, following the notation of [AMS09], we denote the inverse of this matrix by $\mathcal{P}_{\gamma}^{0 \leftarrow \tau}$, where its restriction to $T_{\gamma(\tau)}\mathbb{S}^{n-1}$ is the inverse of the parallel translation operator $\mathcal{P}_{\gamma}^{\tau \leftarrow 0}$.

5.3.2 Key steps towards the proof

Note that for any orthogonal $\boldsymbol{A}_0$, $f\left(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0\right) = f\left(\boldsymbol{A}_0^*\boldsymbol{q}; \boldsymbol{X}_0\right)$. In words, this is the above established fact that the function landscape of $f(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0)$ is a rotated version of that of $f(\boldsymbol{q}; \boldsymbol{X}_0)$. Thus, any local minimizer $\boldsymbol{q}_\star$ of $f(\boldsymbol{q}; \boldsymbol{X}_0)$ is rotated to $\boldsymbol{A}_0\boldsymbol{q}_\star$, one minimizer of $f(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0)$. Also if our algorithm generates iteration sequence $\boldsymbol{q}_0, \boldsymbol{q}_1, \boldsymbol{q}_2, \dots$ for $f(\boldsymbol{q}; \boldsymbol{X}_0)$ upon initialization $\boldsymbol{q}_0$, it will generate the iteration sequence $\boldsymbol{A}_0\boldsymbol{q}_0, \boldsymbol{A}_0\boldsymbol{q}_1, \boldsymbol{A}_0\boldsymbol{q}_2, \dots$ for $f\left(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0\right)$. So w.l.o.g. it is adequate that we prove the convergence results for the case $\boldsymbol{A}_0 = \boldsymbol{I}$. So in this section (Section 5.3), we write $f(\boldsymbol{q})$ to mean $f(\boldsymbol{q}; \boldsymbol{X}_0)$.

We partition the sphere into three regions, for which we label as R_{I}, R_{II}, R_{III}, corresponding to the strongly convex, nonzero gradient, and negative curvature regions, respectively (see Theorem 4.1). That is, R_{I} consists of a union of $2n$ spherical caps of radius $\frac{\mu}{4\sqrt{2}}$, each centered around a signed standard basis vector $\pm\boldsymbol{e}_i$. R_{II} consist of the set difference of a union of $2n$ spherical caps of radius $\frac{1}{20\sqrt{5}}$, centered around the standard basis vectors $\pm\boldsymbol{e}_i$, and R_{I}. Finally, R_{III} covers the rest of the sphere. We say a trust-region step takes an R_{I} step if the current iterate is in R_{I}; similarly for R_{II} and R_{III} steps. Since we use the geometric structures derived in Theorem 4.1 and Corollary 4.2, the conditions

$$\theta \in (0, 1/2), \quad \mu < \min\left\{c_a\theta n^{-1}, c_b n^{-5/4}\right\}, \quad p \geq \frac{C}{\mu^2\theta^2} n^3 \log\frac{n}{\mu\theta} \tag{5.3.7}$$

are always in force.

At each step k of the algorithm, suppose $\boldsymbol{\delta}^{(k)}$ is a minimizer of the trust-region subproblem (5.1.3). We call the step "*constrained*" if $\left\|\boldsymbol{\delta}^{(k)}\right\| = \Delta$ (the minimizer lies on the boundary and hence the constraint is active), and call it "*unconstrained*" if $\|\boldsymbol{\delta}^{(k)}\| < \Delta$ (the minimizer lies in the relative interior and hence the constraint is not in force). Thus, in the unconstrained case the optimality condition is (5.3.5).

The next lemma provides some estimates about ∇f and $\nabla^2 f$ that are useful in various contexts.

Lemma 5.3 *We have the following estimates about ∇f and $\nabla^2 f$:*

$$\begin{aligned} \sup_{\boldsymbol{q} \in \mathbb{S}^{n-1}} \left\| \nabla f\left(\boldsymbol{q}\right) \right\| &\doteq M_{\nabla} \le \sqrt{n} \left\| \boldsymbol{X}_0 \right\|_{\infty}, \\ \sup_{\boldsymbol{q} \in \mathbb{S}^{n-1}} \left\| \nabla^2 f\left(\boldsymbol{q}\right) \right\| &\doteq M_{\nabla^2} \le \frac{n}{\mu} \left\| \boldsymbol{X}_0 \right\|_{\infty}^2, \\ \sup_{\boldsymbol{q}, \boldsymbol{q}' \in \mathbb{S}^{n-1}, \boldsymbol{q} \ne \boldsymbol{q}'} \frac{\left\| \nabla f\left(\boldsymbol{q}\right) - \nabla f\left(\boldsymbol{q}'\right) \right\|}{\left\| \boldsymbol{q} - \boldsymbol{q}' \right\|} &\doteq L_{\nabla} \le \frac{n}{\mu} \left\| \boldsymbol{X}_0 \right\|_{\infty}^2, \\ \sup_{\boldsymbol{q}, \boldsymbol{q}' \in \mathbb{S}^{n-1}, \boldsymbol{q} \ne \boldsymbol{q}'} \frac{\left\| \nabla^2 f\left(\boldsymbol{q}\right) - \nabla^2 f\left(\boldsymbol{q}'\right) \right\|}{\left\| \boldsymbol{q} - \boldsymbol{q}' \right\|} &\doteq L_{\nabla^2} \le \frac{2}{\mu^2} n^{3/2} \left\| \boldsymbol{X}_0 \right\|_{\infty}^3. \end{aligned}$$

Proof See Page 108 under Section 10.1. ■

Our next lemma says if the trust-region step size Δ is small enough, one Riemannian trust-region step reduces the objective value by a certain amount when there is any descent direction.

Lemma 5.4 *Suppose that the trust region size $\Delta \le 1$, and there exists a tangent vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}} \mathbb{S}^{n-1}$ with $\|\boldsymbol{\delta}\| \le \Delta$, such that*

$$f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) \ \le\ f(\boldsymbol{q}) - s$$

for some positive scalar $s \in \mathbb{R}$. Then the trust region subproblem produces a point $\boldsymbol{\delta}_\star$ with

$$f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta}_\star)) \ \le\ f(\boldsymbol{q}) - s + \frac{1}{3}\eta_f \Delta^3,$$

where $\eta_f \doteq M_{\nabla} + 2M_{\nabla^2} + L_{\nabla} + L_{\nabla^2}$ and M_{∇}, M_{∇^2}, L_{∇}, L_{∇^2} are the quantities defined in Lemma 5.3.

Proof See Page 109 under Section 10.2. ■

To show decrease in objective value for R_{II} and R_{III}, now it is enough to exhibit a descent direction for each point in these regions. The next two lemmas help us almost accomplish the goal. For convenience again we choose to state the results for the "canonical" section that is in the vicinity of $\boldsymbol{e}_n$ and the projection map $\boldsymbol{q}\left(\boldsymbol{w}\right) = [\boldsymbol{w}; (1 - \left\|\boldsymbol{w}\right\|^2)^{1/2}]$, with the idea that similar statements hold for other symmetric sections.

Lemma 5.5 *Suppose that the trust region size $\Delta \le 1$, $\boldsymbol{w}^* \nabla g(\boldsymbol{w}) / \left\|\boldsymbol{w}\right\| \ge \beta_g$ for some scalar β_g, and that $\boldsymbol{w}^* \nabla g(\boldsymbol{w}) / \left\|\boldsymbol{w}\right\|$ is L_g-Lipschitz on an open ball $\mathcal{B}\left(\boldsymbol{w}, \frac{3\Delta}{2\pi\sqrt{n}}\right)$ centered at $\boldsymbol{w}$. Then there exists a tangent vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}} \mathbb{S}^{n-1}$ with $\|\boldsymbol{\delta}\| \le \Delta$, such that*

$$f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) \ \le\ f(\boldsymbol{q}) - \min\left\{ \frac{\beta_g^2}{2L_g}, \frac{3\beta_g \Delta}{4\pi\sqrt{n}} \right\}.$$

Proof See Page 110 under Section 10.3. ■

Lemma 5.6 *Suppose that the trust-region size $\Delta \le 1$, $\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w} / \left\|\boldsymbol{w}\right\|^2 \le -\beta_{\frown}$, for some $\beta_{\frown}$, and that $\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w} / \left\|\boldsymbol{w}\right\|^2$ is $L_{\frown}$ Lipschitz on the open ball $\mathcal{B}\left(\boldsymbol{w}, \frac{3\Delta}{2\pi\sqrt{n}}\right)$ centered at $\boldsymbol{w}$. Then there exists a tangent vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ with $\|\boldsymbol{\delta}\| \le \Delta$, such that*

$$f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) \;\le\; f(\boldsymbol{q}) - \min\left\{\frac{2\beta_{\frown}^3}{3L_{\frown}^2}, \frac{3\Delta^2\beta_{\frown}}{8\pi^2 n}\right\}.$$

Proof See Page 111 under Section 10.4. ■

One can take $\beta_g = \beta_{\frown} = c_\star\theta$ as shown in Theorem 4.1, and take the Lipschitz results in Section 4.2 (note that $\left\|\boldsymbol{X}_0\right\|_\infty \le 4\log^{1/2}(np)$ w.h.p. by Lemma 9.11), repeat the argument for other $2n-1$ symmetric regions, and conclude that w.h.p. the objective value decreases by at least a constant amount. The next proposition summarizes the results.

Proposition 5.7 *Assume (5.3.7). In regions R_{II} and R_{III}, each trust-region step reduces the objective value by at least*

$$d_{\mathrm{II}} = \frac{1}{2}\min\left(\frac{c_\star^2 c_a \theta^2 \mu}{n^2\log(np)}, \frac{3\Delta c_\star\theta}{4\pi\sqrt{n}}\right), \quad \text{and} \quad d_{\mathrm{III}} = \frac{1}{2}\min\left(\frac{c_\star^3 c_b \theta^3\mu^4}{n^6\log^3(np)}, \frac{3\Delta^2 c_\star\theta}{8\pi^2 n}\right) \tag{5.3.8}$$

respectively, provided that

$$\Delta < \frac{c_c c_\star \theta\mu^2}{n^{5/2}\log^{3/2}(np)}, \tag{5.3.9}$$

where c_a through c_c are positive constants, and $c_\star$ is as defined in Theorem 4.1.

Proof We only consider the symmetric section in the vicinity of $\boldsymbol{e}_n$ and the claims carry on to others by symmetry. If the current iterate $\boldsymbol{q}^{(k)}$ is in the region R_{II}, by Theorem 4.1, w.h.p., we have $\boldsymbol{w}^* g(\boldsymbol{w}) / \|\boldsymbol{w}\| \ge c_\star\theta$ for the constant $c_\star$. By Proposition 4.12 and Lemma 9.11, w.h.p., $\boldsymbol{w}^* g(\boldsymbol{w}) / \|\boldsymbol{w}\|$ is $C_2 n^2 \log(np)/\mu$-Lipschitz. Therefore, By Lemma 5.4 and Lemma 5.5, a trust-region step decreases the objective value by at least

$$d_{\mathrm{II}} \doteq \min\left(\frac{c_\star^2\theta^2\mu}{2C_2 n^2\log(np)}, \frac{3c_\star\theta\Delta}{4\pi\sqrt{n}}\right) - \frac{c_0 n^{3/2}\log^{3/2}(np)}{3\mu^2}\Delta^3.$$

Similarly, if $\boldsymbol{q}^{(k)}$ is in the region R_{III}, by Proposition 4.11, Theorem 4.1 and Lemma 9.11, w.h.p., $\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w}/\|\boldsymbol{w}\|^2$ is $C_3 n^3\log^{3/2}(np)/\mu^2$-Lipschitz and upper bounded by $-c_\star\theta$. By Lemma 5.4 and Lemma 5.6, a trust-region step decreases the objective value by at least

$$d_{\mathrm{III}} \doteq \min\left(\frac{2c_\star^3\theta^3\mu^4}{3C_3^2 n^6\log^3(np)}, \frac{3\Delta^2 c_\star\theta}{8\pi^2 n}\right) - \frac{c_0 n^{3/2}\log^{3/2}(np)}{3\mu^2}\Delta^3.$$

It can be easily verified that when Δ obeys (5.3.8), (5.3.9) holds. ■

The analysis for R_{I} is slightly trickier. In this region, near each local minimizer, the objective function is strongly convex. So we still expect each trust-region step decreases the objective value. On the other hand, it is very unlikely that we can provide a universal lower bound for the amount of decrease - as the iteration sequence approaches one local minimizer, the movement is expected to be diminishing. Nevertheless, close to the minimizer the trust-region algorithm takes "unconstrained" steps. For constrained R_{I} steps, we will again show reduction in objective value by at least a fixed amount; for unconstrained step, we will show the distance between the iterate and the nearest local minimizer drops down rapidly.

The next lemma concerns the function value reduction for constrained R_{I} steps.

Lemma 5.8 *Suppose the trust-region size $\Delta \le 1$, and that at a given iterate k, $\operatorname{Hess} f\left(\boldsymbol{q}^{(k)}\right) \succeq m_H \mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}}$, and $\left\|\operatorname{Hess} f\left(\boldsymbol{q}^{(k)}\right)\right\| \le M_H$. Further assume the optimal solution $\boldsymbol{\delta}_\star \in T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}$ to the trust-region subproblem (5.1.3) satisfies $\|\boldsymbol{\delta}_\star\| = \Delta$, i.e., the norm constraint is active. Then there exists a tangent vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}$ with $\|\boldsymbol{\delta}\| \le \Delta$, such that*

$$f(\exp_{\boldsymbol{q}^{(k)}}(\boldsymbol{\delta})) \ \le\ f\left(\boldsymbol{q}^{(k)}\right) - \frac{m_H^2\Delta^2}{M_H} + \frac{1}{6}\eta_f\Delta^3,$$

where η_f is defined the same as Lemma 5.4.

Proof See Page 112 under Section 10.5. ■

The next lemma provides an estimate of m_H. Again we will only state the result for the "canonical" section with the "canonical" $\boldsymbol{q}(\boldsymbol{w})$ mapping.

Lemma 5.9 *There exist positive constants C and $c_\sharp$, such that for all $\theta \in (0, 1/2)$ and $\mu < \theta/10$, whenever $p \ge Cn^3 \log \frac{n}{\theta\mu}/(\mu\theta^2)$, it holds with probability at least $1 - \theta\,(np)^{-7} - \exp\left(-0.3\theta np\right) - p^{-10}$ that for all $\boldsymbol{q}$ with $\|\boldsymbol{w}\left(\boldsymbol{q}\right)\| \le \frac{\mu}{4\sqrt{2}}$,*

$$\operatorname{Hess} f\left(\boldsymbol{q}\right) \succeq c_\sharp \frac{\theta}{\mu}\mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}.$$

Proof See Page 113 under Section 10.6. ■

We know that $\|\boldsymbol{X}_0\|_\infty \le 4\log^{1/2}(np)$ w.h.p., and hence by the definition of Riemannian Hessian and Lemma 5.3,

$$M_H \doteq \|\operatorname{Hess} f(\boldsymbol{q})\| \le \left\|\nabla^2 f(\boldsymbol{q})\right\| + \|\nabla f(\boldsymbol{q})\| \le M_{\nabla^2} + M_{\nabla} \le \frac{2n}{\mu}\|\boldsymbol{X}_0\|_\infty^2 \le \frac{16n}{\mu}\log(np),$$

Combining this estimate and Lemma 5.9, and Lemma 5.4, we obtain a concrete lower bound for the reduction of objective value for each constrained R_{I} step.

> **Proposition 5.10** *Assume* (5.3.7)*. Each constrained R_{I} trust-region step (i.e., $\|\boldsymbol{\delta}\| = \Delta$) reduces the objective value by at least*
>
> $$d_{\mathrm{I}} = \frac{c c_\star^2 \theta^2}{\mu n \log(np)} \Delta^2, \tag{5.3.10}$$
>
> *provided*
>
> $$\Delta \le \frac{c' c_\sharp^2 \theta^2 \mu}{n^{5/2} \log^{5/2}(np)}. \tag{5.3.11}$$
>
> *The constant $c_\sharp$ is as defined in Lemma* 5.9 *and c, c' are positive constants.*

Proof We only consider the symmetric section in the vicinity of $\boldsymbol{e}_n$ and the claims carry on to others by symmetry. We have that w.h.p.

$$\left\|\mathrm{Hess}\, f(\boldsymbol{q})\right\| \le \frac{16n}{\mu} \log(np), \quad \text{and} \quad \mathrm{Hess}\, f(\boldsymbol{q}) \succeq c_\sharp \frac{\theta}{\mu} \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}},$$

where $c_\sharp$ is as defined in Lemma 5.9. Combining these estimates with Lemma 5.4 and Lemma 5.8, one trust-region step will find next iterate $\boldsymbol{q}^{(k+1)}$ that decreases the objective value by at least

$$d_{\mathrm{I}} \doteq \frac{c_\sharp^2 \theta^2 / \mu^2}{2n \log\left(np\right) / \mu} \Delta^2 - \frac{c_0 n^{3/2} \log^{3/2}\left(np\right)}{\mu^2} \Delta^3.$$

Finally, by the condition on Δ in (5.3.11) and the assumed conditions (5.3.7), we obtain

$$d_{\mathrm{I}} \ge \frac{c_\sharp^2 \theta^2}{2\mu n \log(np)} \Delta^2 - \frac{c_0 n^{3/2} \log^{3/2}\left(np\right)}{\mu^2} \Delta^3 \ge \frac{c_\sharp^2 \theta^2}{4\mu n \log(np)} \Delta^2,$$

as desired. ■

By the proof strategy for R_{I} we sketched before Lemma 5.8, we expect the iteration sequence ultimately always takes unconstrained steps when it moves very near to a local minimizer. We will show that the following is true: when Δ is small enough, once the iteration sequence starts to take an unconstrained R_{I} step, it will take consecutive unconstrained R_{I} steps afterwards. It takes two steps to show this: (1) upon an unconstrained R_{I} step, the next iterate will stay in R_{I}. It is obvious we can make $\Delta \in O(1)$ to ensure the next iterate stays in $R_{\mathrm{I}} \cup R_{\mathrm{II}}$. To strengthen the result, we use the gradient information. From Theorem 4.1, we expect the magnitudes of the gradients in R_{II} to be lower bounded; on the other hand, in R_{I} where points are near local minimizers, continuity argument implies that the magnitudes of gradients should be upper bounded. We will show that when Δ is small enough, there is a gap between these two bounds, implying the next iterate stays in R_{I}; (2) when Δ is small enough, the step is in fact unconstrained. Again we will only state the result for the "canonical" section with the "canonical" $\boldsymbol{q}(\boldsymbol{w})$ mapping. The next lemma exhibits an absolute lower bound for magnitudes of gradients in R_{II}.

> **Lemma 5.11** *For all $\boldsymbol{q}$ satisfying $\frac{\mu}{4\sqrt{2}} \le \|\boldsymbol{w}(\boldsymbol{q})\| \le \frac{1}{20\sqrt{5}}$, it holds that*
>
> $$\|\mathrm{grad}\, f(\boldsymbol{q})\| \ge \frac{9}{10} \frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|}.$$

Proof See Page 119 under Section 10.7. ■

Assuming (5.3.7), Theorem 4.1 gives that w.h.p. $\boldsymbol{w}^* \nabla g(\boldsymbol{w}) / \|\boldsymbol{w}\| \ge c_\star \theta$. Thus, w.h.p, $\|\mathrm{grad}\, f(\boldsymbol{q})\| \ge 9 c_\star \theta / 10$ for all $\boldsymbol{q} \in R_{\mathrm{II}}$. The next lemma compares the magnitudes of gradients before and after taking an unconstrained R_{I} step. This is crucial to providing upper bound for magnitude of gradient for the next iterate, and also to establishing the ultimate (quadratic) sequence convergence.

> **Lemma 5.12** *Suppose the trust-region size $\Delta \le 1$, and at a given iterate k, $\mathrm{Hess}\, f\left(\boldsymbol{q}^{(k)}\right) \succeq m_H \mathcal{P}_{T_{\boldsymbol{q}^{(k)}} \mathbb{S}^{n-1}}$, and that the unique minimizer $\boldsymbol{\delta}_\star \in T_{\boldsymbol{q}^{(k)}} \mathbb{S}^{n-1}$ to the trust region subproblem (5.1.3) satisfies $\|\boldsymbol{\delta}_\star\| < \Delta$ (i.e., the constraint is inactive). Then, for $\boldsymbol{q}^{(k+1)} = \exp_{\boldsymbol{q}^{(k)}}(\boldsymbol{\delta}_\star)$, we have*
>
> $$\|\,\mathrm{grad}\, f(\boldsymbol{q}^{(k+1)})\| \;\le\; \frac{L_H}{2 m_H^2} \|\,\mathrm{grad}\, f(\boldsymbol{q}^{(k)})\|^2,$$
>
> *where $L_H \doteq \frac{5}{2\mu^2} n^{3/2} \|\boldsymbol{X}_0\|_\infty^3 + \frac{9}{\mu} n \|\boldsymbol{X}_0\|_\infty^2 + 9\sqrt{n} \|\boldsymbol{X}_0\|_\infty$.*

Proof See Page 120 under Section 10.8. ■

We can now bound the Riemannian gradient of the next iterate as

$$\begin{aligned}
\|\,\mathrm{grad}\, f(\boldsymbol{q}^{(k+1)})\| &\le \frac{L_H}{2 m_H^2} \|\,\mathrm{grad}\, f(\boldsymbol{q}^{(k)})\|^2 \\
&\le \frac{L_H}{2 m_H^2} \|[\boldsymbol{U}^*\, \mathrm{Hess}\, f(\boldsymbol{q}^{(k)}) \boldsymbol{U}][\boldsymbol{U}^*\, \mathrm{Hess}\, f(\boldsymbol{q}^{(k)}) \boldsymbol{U}]^{-1}\, \mathrm{grad}\, f(\boldsymbol{q}^{(k)})\|^2 \\
&\le \frac{L_H}{2 m_H^2} \left\|\mathrm{Hess}\, f(\boldsymbol{q}^{(k)})\right\|^2 \Delta^2 = \frac{L_H M_H^2}{2 m_H^2} \Delta^2.
\end{aligned}$$

Obviously, one can make the upper bound small by tuning down Δ. Combining the above lower bound for $\|\mathrm{grad}\, f(\boldsymbol{q})\|$ for $\boldsymbol{q} \in R_{\mathrm{II}}$, one can conclude that when Δ is small, the next iterate $\boldsymbol{q}^{(k+1)}$ stays in R_{I}. Another application of the optimality condition (5.3.5) gives conditions on Δ that guarantees the next trust-region step is also unconstrained. Detailed argument can be found in proof of the following proposition.

> **Proposition 5.13** *Assume (5.3.7). W.h.p, once the trust-region algorithm takes an unconstrained $R_{\mathtt{I}}$ step (i.e., $\|\boldsymbol{\delta}\| < \Delta$), it always takes unconstrained $R_{\mathtt{I}}$ steps, provided that*
>
> $$\Delta \le \frac{c c_\sharp^3 \theta^3 \mu}{n^{7/2} \log^{7/2}(np)}, \tag{5.3.12}$$
>
> *Here c is a positive constant, and $c_\sharp$ is as defined in Lemma 5.9.*

Proof We only consider the symmetric section in the vicinity of $\boldsymbol{e}_n$ and the claims carry on to others by symmetry. Suppose that step k is an unconstrained $R_{\mathtt{I}}$ step. Then

$$\begin{aligned}\|\boldsymbol{w}(\boldsymbol{q}^{(k+1)}) - \boldsymbol{w}(\boldsymbol{q}^{(k)})\| \le \|\boldsymbol{q}^{(k+1)} - \boldsymbol{q}^{(k)}\| &= \| \exp_{\boldsymbol{q}^{(k)}(\boldsymbol{\delta})} - \boldsymbol{q}^{(k)}\| \\ &= \sqrt{2 - 2\cos\|\boldsymbol{\delta}\|} = 2\sin(\|\boldsymbol{\delta}\|/2) \le \|\boldsymbol{\delta}\| < \Delta.\end{aligned}$$

Thus, if $\Delta \le \frac{1}{20\sqrt{5}} - \frac{\mu}{4\sqrt{2}}$, $\boldsymbol{q}^{(k+1)}$ will be in $R_{\mathtt{I}} \cup R_{\mathtt{II}}$. Next, we show that if Δ is sufficiently small, $\boldsymbol{q}^{(k+1)}$ will be indeed in $R_{\mathtt{I}}$. By Lemma 5.12,

$$\begin{aligned}\left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k+1)}\right)\right\| &\le \frac{L_H}{2m_H^2}\left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|^2 \\ &\le \frac{L_H M_H^2}{2m_H^2}\left\|\left[\boldsymbol{U}^* \operatorname{Hess} f\left(\boldsymbol{q}^{(k)}\right)\boldsymbol{U}\right]^{-1}\boldsymbol{U}^* \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|^2 \le \frac{L_H M_H^2}{2m_H^2}\Delta^2,\end{aligned} \tag{5.3.13}$$

where we have used the fact that

$$\left\|\boldsymbol{\delta}^{(k)}\right\| = \left\|\left[\boldsymbol{U}^* \operatorname{Hess} f\left(\boldsymbol{q}^{(k)}\right)\boldsymbol{U}\right]^{-1}\boldsymbol{U}^* \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\| < \Delta,$$

as the step is unconstrained. On the other hand, by Theorem 4.1 and Lemma 5.11, w.h.p.

$$\|\operatorname{grad} f(\boldsymbol{q})\| \ge \beta_{\mathrm{grad}} \doteq \frac{9}{10}c_\star\theta, \quad \forall\, \boldsymbol{q} \in R_{\mathtt{II}}. \tag{5.3.14}$$

Hence, provided

$$\Delta < \frac{m_H}{M_H}\sqrt{\frac{2\beta_{\mathrm{grad}}}{L_H}}, \tag{5.3.15}$$

we have $\boldsymbol{q}^{(k+1)} \in R_{\mathtt{I}}$.

We next show that when Δ is small enough, the next step is also unconstrained. Straight forward calculations give

$$\left\|\boldsymbol{U}\left[\boldsymbol{U}^* \operatorname{Hess} f\left(\boldsymbol{q}^{(k+1)}\right)\boldsymbol{U}\right]^{-1}\boldsymbol{U}^* \operatorname{grad} f\left(\boldsymbol{q}^{(k+1)}\right)\right\| \le \frac{L_H M_H^2}{2m_H^3}\Delta^2.$$

Hence, provided that

$$\Delta < \frac{2m_H^3}{L_H M_H^2}, \tag{5.3.16}$$

we will have

$$\left\| \boldsymbol{U} \left[\boldsymbol{U}^* \operatorname{Hess} f\left(\boldsymbol{q}^{(k+1)}\right) \boldsymbol{U} \right]^{-1} \boldsymbol{U}^* \operatorname{grad} f\left(\boldsymbol{q}^{(k+1)}\right) \right\| < \Delta;$$

in words, the minimizer to the trust-region subproblem for the next step lies in the relative interior of the trust region - the constraint is inactive. By Lemma 5.12 and Lemma 9.11, we have

$$L_H \;=\; C_1 n^{3/2} \log^{3/2}(np)/\mu^2, \tag{5.3.17}$$

w.h.p. for some constant C_1. Combining this and our previous estimates of m_H, M_H, we conclude whenever

$$\Delta \le \min\left\{ \frac{1}{20\sqrt{5}} - \frac{\mu}{4\sqrt{2}}, \frac{c_1 \mu c_\sharp c_\star^{1/2} \theta^{3/2}}{n^{7/4}\log^{7/4}(np)}, \frac{c_2 \mu c_\sharp^3 \theta^3}{n^{7/2}\log^{7/2}(np)} \right\}.$$

for some positive constants c_1 and c_2, w.h.p. our next trust-region step is also an unconstrained R_{I} step. Noting that $c_\star$ and $c_\sharp$ can be made the same by our definition, we make the claimed simplification on Δ. This completes the proof. ■

Finally, we want to show that ultimate unconstrained R_{I} iterates actually converges to one nearby local minimizer rapidly. Lemma 5.12 has established the gradient is diminishing. The next lemma shows the magnitude of gradient serves as a good proxy for distance to the local minimizer.

> **Lemma 5.14** *Let $\boldsymbol{q}_\star \in \mathbb{S}^{n-1}$ such that $\operatorname{grad} f(\boldsymbol{q}_\star) = \boldsymbol{0}$, and $\boldsymbol{\delta} \in T_{\boldsymbol{q}_\star}\mathbb{S}^{n-1}$. Consider a geodesic $\gamma(t) = \exp_{\boldsymbol{q}_\star}(t\boldsymbol{\delta})$, and suppose that on $[0,\tau]$, $\operatorname{Hess} f(\gamma(t)) \succeq m_H \mathcal{P}_{T_{\gamma(t)}\mathbb{S}^{n-1}}$. Then*
>
> $$\|\operatorname{grad} f(\gamma(\tau))\| \;\ge\; m_H \tau \|\boldsymbol{\delta}\|.$$

Proof See Page 120 under Section 10.9. ■

To see this relates the magnitude of gradient to the distance away from the critical point, w.l.o.g., one can assume $\tau = 1$ and consider the point $\boldsymbol{q} = \exp_{\boldsymbol{q}_\star}(\boldsymbol{\delta})$. Then

$$\|\boldsymbol{q}_\star - \boldsymbol{q}\| = \left\|\exp_{\boldsymbol{q}_\star}(\boldsymbol{\delta}) - \boldsymbol{q}\right\| = \sqrt{2 - 2\cos\|\boldsymbol{\delta}\|} = 2\sin(\|\boldsymbol{\delta}\|/2) \le \|\boldsymbol{\delta}\| \le \|\operatorname{grad} f(\boldsymbol{q})\|/m_H,$$

where at the last inequality above we have used Lemma 5.14. Hence, combining this observation with Lemma 5.12, we can derive the asymptotic sequence convergence result as follows.

Proposition 5.15 *Assume (5.3.7) and the conditions in Lemma 5.13. Let $\boldsymbol{q}^{(k_0)} \in R_{\mathrm{I}}$ and the k_0-th step the first unconstrained R_{I} step and $\boldsymbol{q}_\star$ be the unique local minimizer of f over one connected component of R_{I} that contains $\boldsymbol{q}^{(k_0)}$. Then w.h.p., for any positive integer $k' \geq 1$,*

$$\left\| \boldsymbol{q}^{(k_0+k')} - \boldsymbol{q}_\star \right\| \leq \frac{c c_\sharp \theta \mu}{n^{3/2} \log^{3/2}(np)} 2^{-2^{k'}}, \tag{5.3.18}$$

provided that

$$\Delta \leq \frac{c' c_\sharp^2 \theta^2 \mu}{n^{5/2} \log^{5/2}(np)}. \tag{5.3.19}$$

Here $c_\sharp$ is as defined in Lemma 5.9 that can be made equal to $c_s\star$ as defined in Theorem 4.1, and c, c' are positive constants.

Proof By the geometric characterization in Theorem 4.1 and corollary 4.2, f has $2n$ separated local minimizers, each located in R_{I} and within distance $\sqrt{2}\mu/16$ of one of the $2n$ signed basis vectors $\{\pm \boldsymbol{e}_i\}_{i\in[n]}$. Moreover, it is obvious when $\mu \leq 1$, R_{I} consists of $2n$ disjoint connected components. We only consider the symmetric component in the vicinity of $\boldsymbol{e}_n$ and the claims carry on to others by symmetry.

Suppose that k_0 is the index of the first unconstrained iterate in region R_{I}, i.e., $\boldsymbol{q}^{(k_0)} \in R_{\mathrm{I}}$. By Lemma 5.12, for any integer $k' \geq 1$, we have

$$\left\| \mathrm{grad}\, f\left(\boldsymbol{q}^{(k_0+k')}\right) \right\| \leq \frac{2m_H^2}{L_H} \left(\frac{L_H}{2m_H^2} \left\| \mathrm{grad}\, f\left(\boldsymbol{q}^{(k_0)}\right) \right\| \right)^{2^{k'}}. \tag{5.3.20}$$

where L_H is as defined in Lemma 5.12, m_H as the strong convexity parameter for R_{I} defined above.

Now suppose $\boldsymbol{q}_\star$ is the unique local minimizer of f, lies in the same R_{I} component that $q^{(k_0)}$ is located. Let $\gamma_{k'}(t) = \exp_{\boldsymbol{q}_\star}(t\boldsymbol{\delta})$ to be the unique geodesic that connects $\boldsymbol{q}_\star$ and $\boldsymbol{q}^{(k_0+k')}$ with $\gamma_{k'}(0) = \boldsymbol{q}_\star$ and $\gamma_{k'}(1) = \boldsymbol{q}^{(k_0+k')}$. We have

$$\begin{aligned} \left\| \boldsymbol{q}^{(k_0+k')} - \boldsymbol{q}_\star \right\| &\leq \left\| \exp_{\boldsymbol{q}_\star}(\boldsymbol{\delta}) - \boldsymbol{q}_\star \right\| = \sqrt{2 - 2\cos\|\boldsymbol{\delta}\|} = 2\sin(\|\boldsymbol{\delta}\|/2) \\ &\leq \|\boldsymbol{\delta}\| \leq \frac{1}{m_H} \left\| \mathrm{grad}\, f\left(\boldsymbol{q}^{(k_0+k')}\right) \right\| \leq \frac{2m_H}{L_H} \left(\frac{L_H}{2m_H^2} \left\| \mathrm{grad}\, f\left(\boldsymbol{q}^{(k_0)}\right) \right\| \right)^{2^{k'}}, \end{aligned}$$

where at the second line we have repeatedly applied Lemma 5.14.

By the optimality condition (5.3.5) and the fact that $\left\| \boldsymbol{\delta}^{(k_0)} \right\| < \Delta$, we have

$$\frac{L_H}{2m_H^2} \left\| \mathrm{grad}\, f\left(\boldsymbol{q}^{(k_0)}\right) \right\| \leq \frac{L_H}{2m_H^2} M_H \left\| \left[\boldsymbol{U}^* \,\mathrm{Hess}\, f\left(\boldsymbol{q}^{(k_0)}\right) \boldsymbol{U} \right]^{-1} \boldsymbol{U}^* \,\mathrm{grad}\, f\left(\boldsymbol{q}^{(k_0)}\right) \right\| \leq \frac{L_H M_H}{2m_H^2} \Delta.$$

Thus, provided

$$\Delta < \frac{m_H^2}{L_H M_H}, \tag{5.3.21}$$

we can combine the above results and obtain

$$\left\| \boldsymbol{q}^{(k_0+k')} - \boldsymbol{q}_\star \right\| \;\le\; \frac{2m_H}{L_H} 2^{-2^{k'}}.$$

Based on the previous estimates for m_H, M_H and L_H, we obtain that w.h.p.,

$$\left\| \boldsymbol{q}^{(k_0+k')} - \boldsymbol{q}_\star \right\| \;\le\; \frac{c_1 c_\sharp \theta \mu}{n^{3/2} \log^{3/2}(np)} 2^{-2^{k'}}.$$

Moreover, by (5.3.21), w.h.p., it is sufficient to have the trust region size

$$\Delta \le \frac{c_2 c_\sharp^2 \theta^2 \mu}{n^{5/2} \log^{5/2}(np)}.$$

Thus, we complete the proof. ■

Now we are ready to piece together the above technical propositions to prove Theorem 5.1.

Proof [of Theorem 5.1] Assuming (5.3.7) and in addition that

$$\Delta < \min\left\{ \frac{c_1 c_\star \theta \mu^2}{n^{5/2} \log^{3/2}(np)}, \frac{c_2 c_\sharp^3 \theta^3 \mu}{n^{7/2} \log^{7/2}(np)} \right\}$$

for small enough constants c_1 and c_2 and $c_\star$, $c_\sharp$ as defined in Theorem 4.1 and Lemma 5.9 respectively ($c_\star$ and $c_\sharp$ can be set to the same constant value), it can be verified that the conditions of all the above propositions are satisfied.

By the preceding four propositions, a step will either be R_{III}, R_{II}, or constrained R_{I} step that decreases the objective value by at least a certain fixed amount (we call this *Type A*), or be an unconstrained R_{I} step (*Type B*), such that all future steps are unconstrained R_{I} and the sequence converges to one local minimizer quadratically. Hence, regardless the initialization, the whole iteration sequence consists of consecutive Type A steps, followed by consecutive Type B steps. Depending on the initialization, either the Type A phase or the Type B phase can be absent. In any case, from $\boldsymbol{q}^{(0)}$ it takes at most (note $f(\boldsymbol{q}) \ge 0$ always holds)

$$\frac{f\left(\boldsymbol{q}^{(0)}\right)}{\min\left\{d_{\mathrm{I}}, d_{\mathrm{II}}, d_{\mathrm{III}}\right\}} \tag{5.3.22}$$

steps for the iterate sequence to start take consecutive unconstrained R_{I} steps, or to already terminate. In case the iterate sequence continues to take consecutive unconstrained R_{I} steps, Proposition 5.15 implies that it takes at most

$$\log\log\left(\frac{c_5 c_\sharp \theta\mu}{\varepsilon n^{3/2}\log^{3/2}(np)}\right) \tag{5.3.23}$$

steps to obtain an ε-near solution to the $\boldsymbol{q}_\star$ that is contained in the connected subset of $R_{\mathtt{I}}$ that the sequence entered.

Thus, the number of iterations to obtain an ε-near solution to $\boldsymbol{q}_\star$ can be grossly bounded by

$$\begin{aligned} \#\text{Iter} \ &\le \ \frac{f\left(\boldsymbol{q}^{(0)}\right)}{\min\left\{d_{\mathtt{I}}, d_{\mathtt{II}}, d_{\mathtt{III}}\right\}} \ + \ \log\log\left(\frac{c_5 c_\sharp \theta\mu}{\varepsilon n^{3/2}\log^{3/2}(np)}\right) \\ &\le \ \left[\min\left\{\frac{c_3 c_\star^3 \theta^3 \mu^4}{n^6\log^3(np)}, \frac{c_4 c_\sharp^2 \theta^2}{n}\Delta^2\right\}\right]^{-1} f\left(\boldsymbol{q}^{(0)}\right) \ + \ \log\log\left(\frac{c_5 c_\sharp \theta\mu}{\varepsilon n^{3/2}\log^{3/2}(np)}\right), \end{aligned}$$

where we have assumed $p \le \exp(n)$ when comparing the various bounds. Finally, the claimed failure probability comes from a simple union bound with careful bookkeeping. ■

5.4 Extending to convergence for complete dictionaries

Note that for any complete $\boldsymbol{A}_0$ with condition number $\kappa(\boldsymbol{A}_0)$, from Lemma 4.14 we know when p is large enough, w.h.p. one can write the preconditioned $\overline{\boldsymbol{Y}}$ as

$$\overline{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \Xi\boldsymbol{X}_0$$

for a certain Ξ with small magnitude, and $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \texttt{SVD}(\boldsymbol{A}_0)$. Since $\boldsymbol{U}\boldsymbol{V}^*$ is orthogonal,

$$f(\boldsymbol{q}; \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \Xi\boldsymbol{X}_0) = f(\boldsymbol{V}\boldsymbol{U}^*\boldsymbol{q}; \boldsymbol{X}_0 + \boldsymbol{V}\boldsymbol{U}^*\Xi\boldsymbol{X}_0).$$

In words, the function landscape of $f(\boldsymbol{q}; \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \Xi\boldsymbol{X}_0)$ is a rotated version of that of $f(\boldsymbol{q}; \boldsymbol{X}_0 + \boldsymbol{V}\boldsymbol{U}^*\Xi\boldsymbol{X}_0)$. Thus, any local minimizer $\boldsymbol{q}_\star$ of $f(\boldsymbol{q}; \boldsymbol{X}_0 + \boldsymbol{V}\boldsymbol{U}^*\Xi\boldsymbol{X}_0)$ is rotated to $\boldsymbol{U}\boldsymbol{V}^*\boldsymbol{q}_\star$, one minimizer of $f(\boldsymbol{q}; \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \Xi\boldsymbol{X}_0)$. Also if our algorithm generates iteration sequence $\boldsymbol{q}_0, \boldsymbol{q}_1, \boldsymbol{q}_2, \dots$ for $f(\boldsymbol{q}; \boldsymbol{X}_0 + \boldsymbol{V}\boldsymbol{U}^*\Xi\boldsymbol{X}_0)$ upon initialization $\boldsymbol{q}_0$, it will generate the iteration sequence $\boldsymbol{U}\boldsymbol{V}^*\boldsymbol{q}_0, \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{q}_1, \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{q}_2, \dots$ for $f(\boldsymbol{q}; \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \Xi\boldsymbol{X}_0)$. So w.l.o.g. it is adequate that we prove the convergence results for the case $f(\boldsymbol{q}; \boldsymbol{X}_0 + \boldsymbol{V}\boldsymbol{U}^*\Xi\boldsymbol{X}_0)$, corresponding to $\boldsymbol{A}_0 = \boldsymbol{I}$ with perturbation $\widetilde{\Xi} \doteq \boldsymbol{V}\boldsymbol{U}^*\Xi$. So in this section (Section 5.4), we write $f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0})$ to mean $f(\boldsymbol{q}; \boldsymbol{X}_0 + \widetilde{\Xi}\boldsymbol{X}_0)$.

Theorem 4.3 has shown that when

$$\theta \in \left(0, \frac{1}{2}\right), \ \mu \leq \min\left\{\frac{c_a\theta}{n}, \frac{c_b}{n^{5/4}}\right\}, \ p \geq \frac{C}{c_\star^2\theta} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8\left(\boldsymbol{A}_0\right) \log^4\left(\frac{\kappa\left(\boldsymbol{A}_0\right)n}{\mu\theta}\right), \tag{5.4.1}$$

the geometric structure of the landscape is qualitatively unchanged and the $c_\star$ constant can be replaced with $c_\star/2$. Particularly, for this choice of p, Lemma 4.14 implies

$$\|\widetilde{\boldsymbol{\Xi}}\| = \|\boldsymbol{V}\boldsymbol{U}^*\boldsymbol{\Xi}\| \leq \left\|\widetilde{\boldsymbol{\Xi}}\right\| \leq cc_\star\theta\left(\max\left\{\frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu}\right\}\log^{3/2}(np)\right)^{-1} \tag{5.4.2}$$

for a constant c that can be made arbitrarily small by setting the constant C in p sufficiently large. The whole proof is quite similar to that of orthogonal case in the last section. We will only sketch the major changes below. To distinguish with the corresponding quantities in the last section, we use $\widetilde{\cdot}$ to denote the corresponding perturbed quantities here.

- Lemma 5.3: Note that

$$\|\boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\|_\infty \leq \|\boldsymbol{X}_0\|_\infty + \|\widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\|_\infty \leq \|\boldsymbol{X}_0\|_\infty + \sqrt{n}\|\widetilde{\boldsymbol{\Xi}}\|\|\boldsymbol{X}_0\|_\infty \leq 3\|\boldsymbol{X}_0\|_\infty/2,$$

where by (5.4.2) we have used $\|\widetilde{\boldsymbol{\Xi}}\| \leq 1/(2\sqrt{n})$ to simplify the above result. So we obtain

$$\widetilde{M_\nabla} \leq \frac{3}{2}M_\nabla, \ \widetilde{M_{\nabla^2}} \leq \frac{9}{4}M_{\nabla^2}, \ \widetilde{L}_\nabla \leq \frac{9}{4}L_\nabla, \ \widetilde{L}_{\nabla^2} \leq \frac{27}{8}L_{\nabla^2}.$$

- Lemma 5.4: Now we have

$$\widetilde{\eta}_f \doteq \widetilde{M_\nabla} + 2\widetilde{M_{\nabla^2}} + \widetilde{L}_\nabla + \widetilde{L}_{\nabla^2} \leq 4\eta_f.$$

- Lemma 5.5 and Lemma 5.6 are generic and nothing changes.

- Proposition 5.7: We have now $\boldsymbol{w}^*\boldsymbol{g}(\boldsymbol{w})/\|\boldsymbol{w}\| \geq c_\star\theta/2$ by Theorem 4.3 and w.h.p. $\boldsymbol{w}^*\nabla g(\boldsymbol{w})/\|\boldsymbol{w}\|$ is $C_1n^2\log(np)/\mu$-Lipschitz by Proposition 4.12 and the fact $\left\|\boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right\|_\infty \leq 3\|\boldsymbol{X}_0\|_\infty/2$ shown above. Similarly, $\boldsymbol{w}^*\boldsymbol{g}(\boldsymbol{w})/\|\boldsymbol{w}\| \leq -c_\star\theta/2$ by Theorem 4.3 and $\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w}/\|\boldsymbol{w}\|^2$ is $C_2n^3\log^{3/2}(np)/\mu^2$-Lipschitz. Moreover, $\widetilde{\eta}_f \leq 4\eta_f$ as shown above. Since there are only multiplicative constant changes to the various quantities, we conclude

$$\widetilde{d_{\mathrm{II}}} = c_1 d_{\mathrm{II}}, \quad \widetilde{d_{\mathrm{III}}} = c_1 d_{\mathrm{III}} \tag{5.4.3}$$

provided

$$\Delta < \frac{c_2 c_\star \theta \mu^2}{n^{5/2} \log^{3/2}(np)}. \tag{5.4.4}$$

- Lemma 5.8: η_f is changed to $\widetilde{\eta}_f$ with $\widetilde{\eta}_f \le 4\eta_f$ as shown above.

- Lemma 5.9: By (5.3.2), we have

$$\begin{aligned} \left\| \nabla^2 f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla^2 f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| &\le \frac{1}{p} \sum_{k=1}^{p} \left\{ L_{\ddot{h}} \|\widetilde{\boldsymbol{\Xi}}\| \left\|\boldsymbol{x}_k\right\|^2 + \frac{1}{\mu} \left\|\boldsymbol{x}_k \boldsymbol{x}_k^* - \widetilde{\boldsymbol{x}}_k \widetilde{\boldsymbol{x}}_k^*\right\| \right\} \\ &\le \|\widetilde{\boldsymbol{\Xi}}\| \left(L_{\ddot{h}} + 2/\mu + \|\widetilde{\boldsymbol{\Xi}}\|/\mu \right) \sum_{k=1}^{p} \left\|\boldsymbol{x}_k\right\|^2 \le \|\widetilde{\boldsymbol{\Xi}}\| \left(L_{\ddot{h}} + 3/\mu \right) n \left\|\boldsymbol{X}_0\right\|_\infty^2, \end{aligned}$$

where $L_{\ddot{h}}$ is the Lipschitz constant for the function $\ddot{h}_\mu(\cdot)$ and we have used the fact that $\|\widetilde{\boldsymbol{\Xi}}\| \le 1$. Similarly, by 5.3.1,

$$\left\| \nabla f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| \le \frac{1}{p} \sum_{k=1}^{p} \left\{ L_{\dot{h}_\mu} \|\widetilde{\boldsymbol{\Xi}}\| \left\|\boldsymbol{x}_k\right\| + \|\widetilde{\boldsymbol{\Xi}}\| \left\|\boldsymbol{x}_k\right\| \right\} \le \left(L_{\dot{h}_\mu} + 1 \right) \|\widetilde{\boldsymbol{\Xi}}\| \sqrt{n} \left\|\boldsymbol{X}_0\right\|_\infty,$$

where $L_{\dot{h}}$ is the Lipschitz constant for the function $\dot{h}_\mu(\cdot)$. Since $L_{\ddot{h}} \le 2/\mu^2$ and $L_{\dot{h}} \le 1/\mu$, and $\left\|\boldsymbol{X}_0\right\|_\infty \le 4\sqrt{\log(np)}$ w.h.p. (Lemma 9.11). By (5.4.2), w.h.p. we have

$$\left\| \nabla f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| \le \frac{1}{2} c_\sharp \theta, \quad \text{and} \quad \left\| \nabla^2 f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla^2 f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| \le \frac{1}{2} c_\sharp \theta,$$

provided the constant C in (5.4.1) for p is large enough. Thus, by (5.3.4) and the above estimates we have

$$\begin{aligned} \left\| \operatorname{Hess} f(\boldsymbol{q}; \boldsymbol{X}_0) - \operatorname{Hess} f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| &\le \left\| \nabla f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| + \left\| \nabla^2 f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla^2 f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| \\ &\le c_\sharp \theta \le \frac{1}{2} c_\sharp \frac{\theta}{\mu}, \end{aligned}$$

provided $\mu \le 1/2$. So we conclude

$$\operatorname{Hess} f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \succeq \frac{1}{2} c_\sharp \frac{\theta}{\mu} \mathcal{P}_{T_{\boldsymbol{q}}} \mathbb{S}^{n-1} \Longrightarrow \widetilde{m_H} \ge \frac{1}{2} c_\sharp \frac{\theta}{\mu}. \tag{5.4.5}$$

- Proposition 5.10: From the estimate of M_H above Proposition 5.10 and the last point, we have

$$\left\| \operatorname{Hess} f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| \le \frac{36}{\mu} \log(np), \quad \text{and} \quad \operatorname{Hess} f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \succeq \frac{1}{2} c_\sharp \frac{\theta}{\mu} \mathcal{P}_{T_{\boldsymbol{q}}} \mathbb{S}^{n-1}.$$

Also since $\widetilde{\eta}_f \le 4\eta_f$ in Lemma 5.4 and Lemma 5.8, there are only multiplicative constant change to the various quantities. We conclude that

$$\widetilde{d_{\mathrm{I}}} = c_3 d_{\mathrm{I}} \tag{5.4.6}$$

provided that

$$\Delta \le \frac{c_4 c_\sharp^2 \theta^2 \mu}{n^{5/2} \log^{5/2}(np)}. \tag{5.4.7}$$

- Lemma 5.11 is generic and nothing changes.
- Lemma 5.12: $\widetilde{L_H} \le 27 L_H/8$.
- Proposition 5.13: All the quantities involved in determining Δ, m_H, M_H, and L_H, β_{grad} are modified by at most constant multiplicative factors and changed to their respective tilde version, so we conclude that the RTM algorithm always takes unconstrained R_{I} step after taking one, provided that

  $$\Delta \le \frac{c_5 c_\sharp^3 \theta^3 \mu}{n^{7/2} \log^{7/2}(np)}. \tag{5.4.8}$$

- Lemma 5.14:is generic and nothing changes.
- Proposition 5.15: Again m_H, M_H, L_H are changed to $\widetilde{m_H}$, $\widetilde{M_H}$, and $\widetilde{L_H}$, respectively, differing by at most constant multiplicative factors. So we conclude for any integer $k' \ge 1$,

  $$\left\| \boldsymbol{q}^{(k_0+k')} - \boldsymbol{q}_\star \right\| \ \le\ \frac{c_6 c_\sharp \theta \mu}{n^{3/2} \log^{3/2}(np)} 2^{-2^{k'}}, \tag{5.4.9}$$

  provided

  $$\Delta \le \frac{c_7 c_\sharp^2 \theta^2 \mu}{n^{5/2} \log^{5/2}(np)}. \tag{5.4.10}$$

The final proof to Theorem 4.3 is almost identical to that of Theorem 4.1, except for

$$\Delta \le \min\left\{ \frac{c_8 c_\star \theta \mu^2}{n^{5/2} \log^{3/2}(np)}, \frac{c_9 c_\sharp^3 \theta^3 \mu}{n^{7/2} \log^{7/2}(np)} \right\}, \tag{5.4.11}$$

$$\widetilde{\zeta} \doteq \min\left\{ \min_{\boldsymbol{q} \in \overline{R_{\mathrm{II}} \cup R_{\mathrm{III}}}} f\left(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}\right), \max_{\boldsymbol{q} \in \overline{R_{\mathrm{I}}}} f\left(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}\right) \right\}, \tag{5.4.12}$$

and hence all ζ is now changed to $\widetilde{\zeta}$, and also d_{I}, d_{II}, and d_{III} are changed to $\widetilde{d_{\mathrm{I}}}$, $\widetilde{d_{\mathrm{II}}}$, and $\widetilde{d_{\mathrm{III}}}$ as defined above, respectively. The final iteration complexity to each an ε-near solution is hence

$$\#\mathrm{Iter} \le \left[\min\left\{ \frac{c_{10} c_\star^3 \theta^3 \mu^4}{n^6 \log^3(np)}, \frac{c_{11} c_\sharp^2 \theta^2}{n} \Delta^2 \right\} \right]^{-1} f\left(\boldsymbol{q}^{(0)}\right) \ +\ \log\log\left(\frac{c_{12} c_\sharp \theta \mu}{\varepsilon n^{3/2} \log^{3/2}(np)} \right).$$

Hence overall the qualitative behavior of the algorithm is not changed, as compared to that for the orthogonal case. Above c_1 through c_{12} are all positive absolute constants.

Chapter 6

Complete Algorithm Pipeline and Main Results

> An expert problem solver must be endowed with two incompatible qualities – a restless imagination and a patient pertinacity.
>
> Howard W. Eves

For orthogonal dictionaries, from Theorem 4.1 and Corollary 4.2, we know that all the minimizers $\widehat{\boldsymbol{q}}_\star$ are $O(\mu)$ away from their respective nearest "target" $\boldsymbol{q}_\star$, with $\boldsymbol{q}_\star^* \widehat{\boldsymbol{Y}} = \alpha \boldsymbol{e}_i^* \boldsymbol{X}_0$ for certain $\alpha \neq 0$ and $i \in [n]$; in Theorem 5.1, we have shown that w.h.p. the Riemannian TRM algorithm produces a solution $\widehat{\boldsymbol{q}} \in \mathbb{S}^{n-1}$ that is ε away to one of the minimizers, say $\widehat{\boldsymbol{q}}_\star$. Thus, the $\widehat{\boldsymbol{q}}$ returned by the TRM algorithm is $O(\varepsilon + \mu)$ away from $\boldsymbol{q}_\star$. For exact recovery, we use a simple linear programming rounding procedure, which guarantees to exactly produce the optimizer $\boldsymbol{q}_\star$. We then use deflation to sequentially recover other rows of $\boldsymbol{X}_0$. Overall, w.h.p. both the dictionary $\boldsymbol{A}_0$ and sparse coefficient $\boldsymbol{X}_0$ are exactly recovered up to sign permutation, when $\theta \in \Omega(1)$, for orthogonal dictionaries. We summarize relevant technical lemmas and main results in Section 6.1. The same procedure can be used to recover complete dictionaries, though the analysis is slightly more complicated; we present the results in Section 6.2. Our overall algorithmic pipeline for recovering orthogonal dictionaries is sketched as follows.

1. **Estimating one row of $\boldsymbol{X}_0$ by the Riemannian TRM algorithm.** By Theorem 4.1 (resp. Theorem 4.3) and Theorem 5.1 (resp. Theorem 5.2), starting from any $\boldsymbol{q}^{(0)} \in \mathbb{S}^{n-1}$, when the relevant parameters are set appropriately (say as $\mu_\star$ and $\Delta_\star$), w.h.p., our Riemannian TRM algorithm finds a local

minimizer $\widehat{\boldsymbol{q}}$, with $\boldsymbol{q}_\star$ the nearest target that exactly recovers one row of $\boldsymbol{X}_0$ and $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \in O(\mu)$ (by setting the target accuracy of the TRM as, say, $\varepsilon = \mu$).

2. **Recovering one row of $\boldsymbol{X}_0$ by rounding.** To obtain the target solution $\boldsymbol{q}_\star$ and hence recover (up to scale) one row of $\boldsymbol{X}_0$, we solve the following linear program:

$$\text{minimize}_{\boldsymbol{q}} \left\|\boldsymbol{q}^* \widehat{\boldsymbol{Y}}\right\|_1, \quad \text{subject to} \quad \langle \boldsymbol{r}, \boldsymbol{q}\rangle = 1, \tag{6.0.1}$$

with $\boldsymbol{r} = \widehat{\boldsymbol{q}}$. We show in Lemma 6.2 (resp. Lemma 6.4) that when $\langle \widehat{\boldsymbol{q}}, \boldsymbol{q}_\star\rangle$ is sufficiently large, implied by μ being sufficiently small, w.h.p. the minimizer of (6.0.1) is exactly $\boldsymbol{q}_\star$, and hence one row of $\boldsymbol{X}_0$ is recovered by $\boldsymbol{q}_\star^* \widehat{\boldsymbol{Y}}$.

3. **Recovering all rows of $\boldsymbol{X}_0$ by deflation.** Once ℓ rows of $\boldsymbol{X}_0$ ($1 \le \ell \le n-2$) have been recovered, say, by unit vectors $\boldsymbol{q}_\star^1, \dots, \boldsymbol{q}_\star^\ell$, one takes an orthonormal basis $\boldsymbol{U}$ for $[\operatorname{span}\left(\boldsymbol{q}_\star^1, \dots, \boldsymbol{q}_\star^\ell\right)]^\perp$, and minimizes the new function $h(\boldsymbol{z}) \doteq f(\boldsymbol{U}\boldsymbol{z}; \widehat{\boldsymbol{Y}})$ on the sphere $\mathbb{S}^{n-\ell-1}$ with the Riemannian TRM algorithm (though conservative, one can again set parameters as $\mu_\star$, $\Delta_\star$, as in Step 1) to produce a $\widehat{\boldsymbol{z}}$. Another row of $\boldsymbol{X}_0$ is then recovered via the LP rounding (6.0.1) with input $\boldsymbol{r} = \boldsymbol{U}\widehat{\boldsymbol{z}}$ (to produce $\boldsymbol{q}_\star^{\ell+1}$). Finally, by repeating the procedure until depletion, one can recover all the rows of $\boldsymbol{X}_0$.

4. **Reconstructing the dictionary $\boldsymbol{A}_0$.** By solving the linear system $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}_0$, one can obtain the dictionary $\boldsymbol{A}_0 = \boldsymbol{Y}\boldsymbol{X}_0^* \left(\boldsymbol{X}_0 \boldsymbol{X}_0^*\right)^{-1}$.

6.1 Recovering orthogonal dictionaries

Theorem 6.1 (Main theorem - recovering orthogonal dictionaries) *Assume the dictionary $\boldsymbol{A}_0$ is orthogonal and we take $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}$. Suppose $\theta \in (0, 1/3)$, $\mu_\star < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}$, and $p \ge C n^3 \log \frac{n}{\mu_\star \theta} / \left(\mu_\star^2 \theta^2\right)$. The above algorithmic pipeline with parameter setting*

$$\Delta_\star \le \min\left\{ \frac{c_c c_\star \theta \mu_\star^2}{n^{5/2} \log^{5/2}(np)}, \frac{c_d c_\star^3 \theta^3 \mu_\star}{n^{7/2} \log^{7/2}(np)} \right\}, \tag{6.1.1}$$

recovers the dictionary $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ in polynomial time, with failure probability bounded by $c_e p^{-6}$. Here $c_\star$ is as defined in Theorem 4.1, and c_a through c_e, and C are all positive constants.

Towards a proof of the above theorem, it remains to be shown the correctness of the rounding and deflation procedures.

Proof of LP rounding. The following lemma shows w.h.p. the rounding will return the desired $\boldsymbol{q}_\star$, provided the estimated $\widehat{\boldsymbol{q}}$ is already near to it.

> **Lemma 6.2 (LP rounding - orthogonal dictionary)** *There exists a positive constant C, such that for all $\theta \in (0, 1/3)$, and $p \geq Cn^2 \log(n/\theta)/\theta$, with probability at least $1 - 2p^{-10} - \theta(n-1)^{-7}p^{-7} - \exp\left(-0.3\theta(n-1)p\right)$, the rounding procedure (6.0.1) returns $\boldsymbol{q}_\star$ for any input vector $\boldsymbol{r}$ that satisfies*
>
> $$\langle \boldsymbol{r}, \boldsymbol{q}_\star \rangle \geq 249/250.$$

Proof See Page 124 under Section 11.1. ■

Since $\langle \widehat{\boldsymbol{q}}, \boldsymbol{q}_\star \rangle = 1 - \|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\|^2/2$, and $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \in O(\mu)$, it is sufficient when μ is smaller than some small constant.

Proof sketch of deflation. We show the deflation works by induction. To understand the deflation procedure, it is important to keep in mind that the "target" solutions $\left\{\boldsymbol{q}_\star^i\right\}_{i=1}^n$ are orthogonal to each other. W.l.o.g., suppose we have found the first ℓ unit vectors $\boldsymbol{q}_\star^1, \dots, \boldsymbol{q}_\star^\ell$ which recover the first ℓ rows of $\boldsymbol{X}_0$. Correspondingly, we partition the target dictionary $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ as

$$\boldsymbol{A}_0 = [\boldsymbol{V}, \boldsymbol{V}^\perp], \quad \boldsymbol{X}_0 = \begin{bmatrix} \boldsymbol{X}_0^{[\ell]} \\ \boldsymbol{X}_0^{[n-\ell]} \end{bmatrix}, \tag{6.1.2}$$

where $\boldsymbol{V} \in \mathbb{R}^{n \times \ell}$, and $\boldsymbol{X}_0^{[\ell]} \in \mathbb{R}^{\ell \times n}$ denotes the submatrix with the first ℓ rows of $\boldsymbol{X}_0$. Let us define a function: $f_{n-\ell}^{\downarrow} : \mathbb{R}^{n-\ell} \mapsto \mathbb{R}$ by

$$f_{n-\ell}^{\downarrow}(\boldsymbol{z}; \boldsymbol{W}) \doteq \frac{1}{p} \sum_{k=1}^{p} h_\mu(\boldsymbol{z}^* \boldsymbol{w}_k), \tag{6.1.3}$$

for any matrix $\boldsymbol{W} \in \mathbb{R}^{(n-\ell) \times p}$. Then by (3.3.2), our objective function is equivalent to

$$h(\boldsymbol{z}) = f(\boldsymbol{U}\boldsymbol{z}; \boldsymbol{A}_0\boldsymbol{X}_0) = f_{n-\ell}^{\downarrow}(\boldsymbol{z}; \boldsymbol{U}^*\boldsymbol{A}_0\boldsymbol{X}_0) = f_{n-\ell}^{\downarrow}(\boldsymbol{z}; \boldsymbol{U}^*\boldsymbol{V}\boldsymbol{X}_0^{[\ell]} + \boldsymbol{U}^*\boldsymbol{V}^\perp\boldsymbol{X}_0^{[n-\ell]}).$$

Since the columns of the orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{n \times (n-\ell)}$ forms the orthogonal complement of span $\left(\boldsymbol{q}_\star^1, \cdots, \boldsymbol{q}_\star^\ell\right)$, it is obvious that $\boldsymbol{U}^*\boldsymbol{V} = \boldsymbol{0}$. Therefore, we obtain

$$h(\boldsymbol{z}) = f_{n-\ell}^{\downarrow}(\boldsymbol{z}; \boldsymbol{U}^*\boldsymbol{V}^\perp\boldsymbol{X}_0^{[n-\ell]}).$$

Since $\boldsymbol{U}^*\boldsymbol{V}^\perp$ is orthogonal and $\boldsymbol{X}_0^{[n-\ell]} \sim_{i.i.d.} \mathrm{BG}(\theta)$, this is another instance of orthogonal dictionary learning problem with reduced dimension. If we keep the parameter settings $\mu_\star$ and $\Delta_\star$ as Theorem 6.1, the conditions of Theorem 4.1 and Theorem 5.1 for all cases with reduced dimensions are still valid. So w.h.p., the TRM algorithm returns a $\widehat{\boldsymbol{z}}$ such that $\|\widehat{\boldsymbol{z}} - \boldsymbol{z}_\star\| \in O(\mu_\star)$ where $\boldsymbol{z}_\star$ is a "target" solution that recovers a row of $\boldsymbol{X}_0$:

$$\boldsymbol{z}_\star^* \boldsymbol{U}^* \boldsymbol{V}^\perp \boldsymbol{X}_0^{[n-\ell]} = \boldsymbol{z}_\star^* \boldsymbol{U}^* \boldsymbol{A}_0 \boldsymbol{X}_0 = \alpha \boldsymbol{e}_i^* \boldsymbol{X}_0, \quad \text{for some } i \notin [\ell].$$

So pulling everything back in the original space, the effective target is $\boldsymbol{q}_\star^{\ell+1} \doteq \boldsymbol{U}\boldsymbol{z}_\star$, and $\boldsymbol{U}\widehat{\boldsymbol{z}}$ is our estimation obtained from the TRM algorithm. Moreover,

$$\|\boldsymbol{U}\widehat{\boldsymbol{z}} - \boldsymbol{U}\boldsymbol{z}_\star\| = \|\widehat{\boldsymbol{z}} - \boldsymbol{z}_\star\| \in O(\mu_\star).$$

Thus, by Lemma 6.2, one successfully recovers $\boldsymbol{U}\boldsymbol{z}_\star$ from $\boldsymbol{U}\widehat{\boldsymbol{z}}$ w.h.p. when $\mu_\star$ is smaller than a constant. The overall failure probability can be obtained via a simple union bound and simplification of the exponential tails with inverse polynomials in p.

6.2 Recovering complete dictionaries

By working with the preconditioned data samples $\widehat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} \doteq \sqrt{\theta p}\left(\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}$,[1] we can use a similar procedure described above to recover complete dictionaries.

> **Theorem 6.3 (Main theorem - recovering complete dictionaries)** *Assume the dictionary $\boldsymbol{A}_0$ is complete with condition number $\kappa(\boldsymbol{A}_0)$ and we take $\widehat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}}$. Suppose $\theta \in (0, 1/3)$, $\mu_\star < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}$, and $p \ge \frac{C}{c_\star^2 \theta} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8(\boldsymbol{A}_0) \log^4\left(\frac{\kappa(\boldsymbol{A}_0) n}{\mu\theta}\right)$. The algorithmic pipeline with parameter setting*
>
> $$\Delta_\star \le \min\left\{\frac{c_c c_\star \theta \mu_\star^2}{n^{5/2}\log^{5/2}(np)}, \frac{c_d c_\star^3 \theta^3 \mu_\star}{n^{7/2}\log^{7/2}(np)}\right\}, \tag{6.2.1}$$
>
> *recovers the dictionary $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ in polynomial time, with failure probability bounded by $c_e p^{-6}$. Here $c_\star$ is as defined in Theorem 4.1, and c_a through c_f, and C are all positive constants.*

Similar to the orthogonal case, we need to show the correctness of the rounding and deflation procedures so that the theorem above holds.

Proof of LP rounding The result of the LP rounding is only slightly different from that of the orthogonal case in Lemma 6.2, so is the proof.

[1] In practice, the parameter θ might not be know beforehand. However, because it only scales the problem, it does not affect the overall qualitative aspect of results.

> **Lemma 6.4 (LP rounding - complete dictionary)** *There exists a positive constant C, such that for all $\theta \in (0, 1/3)$, and $p \geq \frac{C}{c_\star^2 \theta} \max \left\{ \frac{n^4}{\mu^4}, \frac{n^5}{\mu^2} \right\} \kappa^8 \left(\boldsymbol{A}_0 \right) \log^4 \left(\frac{\kappa(\boldsymbol{A}_0) n}{\mu \theta} \right)$, with probability at least $1 - 3p^{-8} - \theta(n-1)^{-7} p^{-7} - \exp\left(-0.3\theta(n-1)p \right)$, the rounding procedure (6.0.1) returns $\boldsymbol{q}_\star$ for any input vector $\boldsymbol{r}$ that satisfies*
>
> $$\langle \boldsymbol{r}, \boldsymbol{q}_\star \rangle \geq 249/250.$$

Proof See Page 126 under Section 11.2. ■

Proof sketch of deflation. We use a similar induction argument as for the orthogonal case to show the deflation works. Compared to the orthogonal case, the tricky part here is that the target vectors $\left\{ \boldsymbol{q}_\star^i \right\}_{i=1}^n$ are not necessarily orthogonal to each other, but they are almost so. W.l.o.g., let us again assume that $\boldsymbol{q}_\star^1, \dots, \boldsymbol{q}_\star^\ell$ recover the first ℓ rows of $\boldsymbol{X}_0$, and similarly partition the matrix $\boldsymbol{X}_0$ as in (6.1.2).

By Lemma 4.14 and (4.3.2), we can write $\overline{\boldsymbol{Y}} = (\boldsymbol{Q} + \boldsymbol{\Xi}) \boldsymbol{X}_0$ for some orthogonal matrix $\boldsymbol{Q}$ and small perturbation $\boldsymbol{\Xi}$ with $\|\boldsymbol{\Xi}\| \leq \delta < 1/10$ for some large p as usual. Similar to the orthogonal case, we have

$$h(\boldsymbol{z}) = f(\boldsymbol{U}\boldsymbol{z}; (\boldsymbol{Q} + \boldsymbol{\Xi})\boldsymbol{X}_0) = f_{n-\ell}^{\downarrow}(\boldsymbol{z}; \boldsymbol{U}^*(\boldsymbol{Q} + \boldsymbol{\Xi})\boldsymbol{X}_0),$$

where $f_{n-\ell}^{\downarrow}$ is defined the same as in (6.1.3). Next, we show that the matrix $\boldsymbol{U}^*(\boldsymbol{Q} + \boldsymbol{\Xi})\boldsymbol{X}_0$ can be decomposed as $\boldsymbol{U}^* \boldsymbol{V} \boldsymbol{X}_0^{[n-\ell]} + \boldsymbol{\Delta}$, where $\boldsymbol{V} \in \mathbb{R}^{(n-\ell) \times n}$ is orthogonal and $\boldsymbol{\Delta}$ is a small perturbation matrix. More specifically, we show that

> **Lemma 6.5** *Suppose the matrices $\boldsymbol{U} \in \mathbb{R}^{n \times (n-\ell)}$, $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ are orthogonal as defined above, $\boldsymbol{\Xi}$ is a perturbation matrix with $\|\boldsymbol{\Xi}\| \leq 1/20$, then*
>
> $$\boldsymbol{U}^* \left(\boldsymbol{Q} + \boldsymbol{\Xi} \right) \boldsymbol{X}_0 = \boldsymbol{U}^* \boldsymbol{V} \boldsymbol{X}_0^{[n-\ell]} + \boldsymbol{\Delta}, \tag{6.2.2}$$
>
> *where $\boldsymbol{V} \in \mathbb{R}^{n \times (n-\ell)}$ is a orthogonal matrix spans the same subspace as that of $\boldsymbol{U}$, and the norms of $\boldsymbol{\Delta}$ is bounded by*
>
> $$\|\boldsymbol{\Delta}\|_{\ell^1 \to \ell^2} \leq 16\sqrt{n} \|\boldsymbol{\Xi}\| \|\boldsymbol{X}_0\|_\infty, \quad \|\boldsymbol{\Delta}\| \leq 16 \|\boldsymbol{\Xi}\| \|\boldsymbol{X}_0\|, \tag{6.2.3}$$
>
> *where $\|\boldsymbol{W}\|_{\ell^1 \to \ell^2} = \sup_{\|\boldsymbol{z}\|_1 = 1} \|\boldsymbol{W}\boldsymbol{z}\| = \max_k \|\boldsymbol{w}_k\|$ denotes the max column ℓ^2-norm of a matrix $\boldsymbol{W}$.*

Proof See Page 126 under Section 11.3. ■

Since $\boldsymbol{U}\boldsymbol{V}$ is orthogonal and $\boldsymbol{X}_0^{[n-\ell]} \sim_{i.i.d.} \mathrm{BG}(\theta)$, we come into another instance of perturbed dictionary learning problem with reduced dimension

$$h(\boldsymbol{z}) = f_{n-\ell}^{\downarrow}(\boldsymbol{z}; \boldsymbol{U}^*\boldsymbol{V}\boldsymbol{X}_0^{[n-\ell]} + \boldsymbol{\Delta}).$$

Since our perturbation analysis in proving Theorem 4.3 and Theorem 5.2 solely relies on the fact that $\|\boldsymbol{\Delta}\|_{\ell^1\to\ell^2} \le C\|\boldsymbol{\Xi}\|\sqrt{n}\|\boldsymbol{X}_0\|_\infty$, it is enough to make p large enough so that the theorems are still applicable for the reduced version $f_{n-\ell}^{\downarrow}(\boldsymbol{z}; \boldsymbol{U}^*\boldsymbol{V}\boldsymbol{X}_0^{[n-\ell]} + \boldsymbol{\Delta})$. Thus, by invoking Theorem 4.3 and Theorem 5.2, the TRM algorithm provably returns one $\widehat{\boldsymbol{z}}$ such that $\widehat{\boldsymbol{z}}$ is near to a perturbed optimal $\widehat{\boldsymbol{z}}_\star$ with

$$\widehat{\boldsymbol{z}}_\star^*\boldsymbol{U}^*\boldsymbol{V}\boldsymbol{X}_0^{[n-\ell]} = \boldsymbol{z}_\star^*\boldsymbol{U}^*\boldsymbol{V}\boldsymbol{X}_0^{[n-\ell]} + \boldsymbol{z}_\star^*\boldsymbol{\Delta} = \alpha\boldsymbol{e}_i^*\boldsymbol{X}_0, \quad \text{for some } i \notin [\ell], \tag{6.2.4}$$

where $\boldsymbol{z}_\star$ with $\|\boldsymbol{z}_\star\| = 1$ is the exact solution. More specifically, Corollary 4.4 implies

$$\|\widehat{\boldsymbol{z}} - \widehat{\boldsymbol{z}}_\star\| \le \sqrt{2}\mu_\star/7.$$

Next, we show that $\widehat{\boldsymbol{z}}$ is also very near to the exact solution $\boldsymbol{z}_\star$. Indeed, the identity (6.2.4) implies

$$\begin{aligned} &(\widehat{\boldsymbol{z}}_\star - \boldsymbol{z}_\star)^* \boldsymbol{U}^*\boldsymbol{V}\boldsymbol{X}_0^{[n-\ell]} = \boldsymbol{z}_\star^*\boldsymbol{\Delta} \\ \Longrightarrow\ & \widehat{\boldsymbol{z}}_\star - \boldsymbol{z}_\star = \left[(\boldsymbol{X}_0^{[n-\ell]})^*\boldsymbol{V}^*\boldsymbol{U}\right]^\dagger \boldsymbol{\Delta}^*\boldsymbol{z}_\star = \boldsymbol{U}^*\boldsymbol{V}\left[(\boldsymbol{X}_0^{[n-\ell]})^*\right]^\dagger \boldsymbol{\Delta}^*\boldsymbol{z}_\star \end{aligned} \tag{6.2.5}$$

where $\boldsymbol{W}^\dagger = (\boldsymbol{W}^*\boldsymbol{W})^{-1}\boldsymbol{W}^*$ denotes the pseudo inverse of a matrix $\boldsymbol{W}$ with full column rank. Hence, by (6.2.5) we can bound the distance between $\widehat{\boldsymbol{z}}_\star$ and $\boldsymbol{z}_\star$ by

$$\|\widehat{\boldsymbol{z}}_\star - \boldsymbol{z}_\star\| \le \left\|\left[(\boldsymbol{X}_0^{[n-\ell]})^*\right]^\dagger\right\| \|\boldsymbol{\Delta}\| \le \sigma_{\min}^{-1}(\boldsymbol{X}_0^{[n-\ell]})\|\boldsymbol{\Delta}\|$$

By Lemma B.12, when $p \ge \Omega(n^2\log n)$, w.h.p.,

$$\theta p/2 \le \sigma_{\min}(\boldsymbol{X}_0^{[n-\ell]}(\boldsymbol{X}_0^{[n-\ell]})^*) \le \left\|\boldsymbol{X}_0^{[n-\ell]}(\boldsymbol{X}_0^{[n-\ell]})^*\right\| \le \|\boldsymbol{X}_0\boldsymbol{X}_0^*\| \le 3\theta p/2.$$

Hence, combined with Lemma 6.5, we obtain

$$\sigma_{\min}^{-1}(\boldsymbol{X}_0^{[n-\ell]}) \le \sqrt{\frac{2}{\theta p}}, \quad \|\boldsymbol{\Delta}\| \le 28\sqrt{\theta p}\|\boldsymbol{\Xi}\|/\sqrt{2},$$

which implies that $\|\widehat{\boldsymbol{z}}_\star - \boldsymbol{z}_\star\| \le 28\|\boldsymbol{\Xi}\|$. Thus, combining the results above, we obtain

$$\|\widehat{\boldsymbol{z}} - \boldsymbol{z}_\star\| \le \|\widehat{\boldsymbol{z}} - \widehat{\boldsymbol{z}}_\star\| + \|\widehat{\boldsymbol{z}}_\star - \boldsymbol{z}_\star\| \le \sqrt{2}\mu_\star/7 + 28\|\boldsymbol{\Xi}\|.$$

Lemma 4.14, and in particular (4.3.2), for our choice of p as in Theorem 4.3, $\|\boldsymbol{\Xi}\| \le c\mu_\star^2 n^{-3/2}$, where c can be made smaller by making the constant in p larger. For $\mu_\star$ sufficiently small, we conclude that

$$\|\boldsymbol{U}\widehat{\boldsymbol{z}} - \boldsymbol{U}\boldsymbol{z}_\star\| = \|\widehat{\boldsymbol{z}} - \boldsymbol{z}_\star\| \leq 2\mu_\star/7.$$

In words, the TRM algorithm returns a $\widehat{\boldsymbol{z}}$ such that $\boldsymbol{U}\widehat{\boldsymbol{z}}$ is very near to one of the unit vectors $\left\{\boldsymbol{q}_\star^i\right\}_{i=1}^n$, such that $(\boldsymbol{q}_\star^i)^*\overline{\boldsymbol{Y}} = \alpha \boldsymbol{e}_i^* \boldsymbol{X}_0$ for some $\alpha \neq 0$. For $\mu_\star$ smaller than a fixed constant, one will have

$$\left\langle \boldsymbol{U}\widehat{\boldsymbol{z}}, \boldsymbol{q}_\star^i \right\rangle \geq 249/250,$$

and hence by Lemma 6.4, the LP rounding exactly returns the optimal solution $\boldsymbol{q}_\star^i$ upon the input $\boldsymbol{U}\widehat{\boldsymbol{z}}$.

The proof sketch above explains why the recursive TRM plus rounding works. The overall failure probability can be obtained via a simple union bound and simplifications of the exponential tails with inverse polynomials in p.

Chapter 7

Numerical Simulations

> The first principle is that you must not fool yourself – and you are the easiest person to fool.
>
> Richard Feynman

To corroborate our theory, we experiment with dictionary recovery on simulated data. For simplicity, we focus on recovering orthogonal dictionaries and we declare success once a single row of the coefficient matrix is recovered. The implementation is based on the Manopt package [BMAS14] with modifications described in Section 2.6.

For the same $\boldsymbol{X}_0$, function landscapes of general orthogonal $\boldsymbol{A}_0$ are exactly rotated versions of that of $\boldsymbol{A}_0 = \boldsymbol{I}$. Thus, w.l.o.g. we set the dictionary as $\boldsymbol{A}_0 = \boldsymbol{I} \in \mathbb{R}^{n \times n}$. We fix $p = 5n^3$ and $5n^2 \log n$ respectively, and each column of the coefficient matrix $\boldsymbol{X}_0 \in \mathbb{R}^{n \times p}$ has exactly k nonzero entries, chosen uniformly random from $\binom{[n]}{k}$. These nonzero entries are i.i.d. standard normals. This is slightly different from the Bernoulli-Gaussian model we assumed for analysis. For n reasonably large, these two models produce similar behavior. For the sparsity surrogate defined in (3.3.3), we fix the parameter $\mu = 10^{-2}$.

To see how the admissible sparsity level varies with the dimension, which our theory primarily is about, we vary the dictionary dimension n and the sparsity k both between 1 and 120;[1] for every pair of (k, n), we randomly generated $T = 5$ instances of $\boldsymbol{X}_0$ and run the TRM algorithm independently from random initializations. Because the optimal solutions are signed coordinate vectors $\{\boldsymbol{e}_i\}_{i=1}^n$, for a solution $\widehat{\boldsymbol{q}}$ returned

[1] Storage is the bottleneck here, as for each instance an almost dense $n \times 5n^3$ matrix needs to be stored. On the other hand, in typical applications of DL the dimension n is normally not significantly larger than 100.

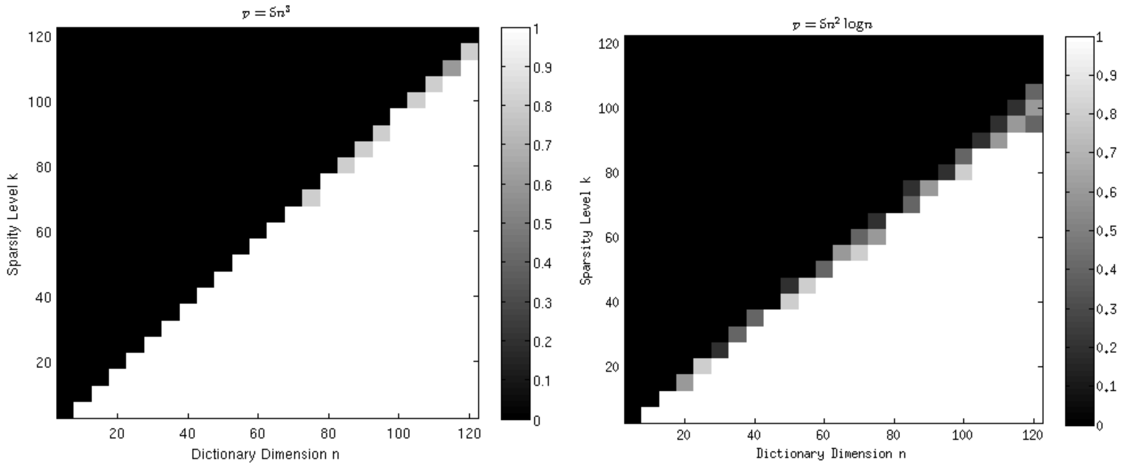


Figure 7.1: Phase transition for recovering a single sparse vector under the dictionary learning model with the sample complexities $p = 5n^3$ and $p = 5n^2 \log n$.

by the TRM algorithm, we define the reconstruction error (RE) to be

$$\mathtt{RE} = \min_{i \in [n]} \left(\|\widehat{\boldsymbol{q}} - \boldsymbol{e}_i\| , \|\widehat{\boldsymbol{q}} + \boldsymbol{e}_i\| \right). \tag{7.0.1}$$

The trial is determined to be a success once $\mathtt{RE} \le \mu$, with the idea that this indicates $\widehat{\boldsymbol{q}}$ is already very near the target and the target can likely be recovered via the LP rounding we described (which we do not implement here). Figure 7.1 shows the phase transition in the (n, k) plane for the orthogonal case. It is obvious that our TRM algorithm can work well into the linear region in this setting whenever $p \in O(n^3)$, perhaps even when $p \in O(n^2 \log n)$. The caveat is that the TRM algorithm was randomly initialized, whereas our results allow arbitrary initializations, which might require slightly higher sample complexity.

To see the sample complexity to ensure the finite-sample convergence (Theorem 4.1), which implies arbitrary initializations lead to “recovery” (up to μ error), we fix $n = 50, 100, 150$, and $p = 5n \log n, 5n^2 \log n, 5n^3$, respectively. For each (n, p) pair, we fixed the sparsity level as $n/2$ and randomly generated an $\boldsymbol{X}_0$, fixed the instance and tested for recovery by the above described criterion with $R = 20$ independent random initializations. The empirical success ratio is calculated. Intuitively, success ratio being one is a strong indicator that the finite-sample function landscape already has the benign $\mathcal{X}$-ness shape. The result is as follows: for $p = 5n^2 \log n$ and $p = 5n^3$, the success ratios are always one, while for $p = 5n \log n$, the ratio is $0.15, 0, 0$ respectively for different n’s. So $O(n^2 \log n)$ samples may already suffice to ensure that the finite-sample function landscape lie in $\mathcal{X}$ family, with the caveat that we only looked at fixed problem instances with a single sparsity level.

Chapter 8

Discussion

Prediction is very difficult, especially if it's about the future.

Niels Bohr

For recovery of complete dictionaries, the LP program approach in [SWW12] that works with $\theta \le O(1/\sqrt{n})$ only demands $p \ge \Omega(n^2 \log^2 n)$, which has recently been reduced to $O(n \log n)$ [Ada16, BN16] (see also [LV15]), matching the lower bound $\Omega(n \log n)$ (i.e., when $\theta \sim 1/n$). The sample complexity stated in Theorem 6.3 is obviously much higher. It is interesting to see whether such growth in complexity is intrinsic to working in the linear regime. Though our experiments seemed to suggest the necessity of $p \sim O(n^3)$ or possibly $O(n^2 \log n)$ for the orthogonal case, there could be other efficient algorithms that demand much less. Tweaking the following three points will likely improve the complexity: (1) the ℓ^1 proxy. The derivatives of the $\log\cosh$ function we adopted entail the $\tanh$ function, which is not amenable to effective approximation and affects the obtained sample complexity; (2) geometric characterization and algorithm analysis. It seems working directly on the sphere (i.e., in the $\boldsymbol{q}$ space) could simplify and possibly improve certain parts of the analysis; (3) treating the complete case directly, rather than using (pessimistic) bounds to treat it as a perturbation of the orthogonal case. Particularly, general linear transforms may change the space significantly, such that preconditioning and comparing to the orthogonal transforms may not be the most efficient way to proceed.

It is possible to extend the current analysis to other dictionary settings. Our geometric structures and algorithms allow plug-and-play noise analysis. Nevertheless, we believe a more stable way of dealing with noise is to directly extract the whole dictionary, i.e., to consider geometry and optimization (and perturbation) over the orthogonal group. This will require additional nontrivial technical work, but likely feasible thanks to availability of relatively complete knowledge of the orthogonal group [EAS98, AMS09]. A substantial leap forward would be to extend the methodology to recovery of *structured* overcomplete dictionaries, such as tight frames. Though there is no natural elimination of one variable, one can consider the marginalization of the objective function w.r.t. the coefficients and work with the resulting implicit function [1], or possibly work directly in the product space. For the coefficient model, as we alluded to in Section 3.4, our analysis and results likely can be carried through to coefficients with statistical dependence and physical constraints.

The connection to ICA we discussed in Section 3.4 suggests our geometric characterization and algorithms can be modified for the ICA problem. This likely will provide new theoretical insights and computational schemes to ICA. In the surge of theoretical understanding of nonconvex heuristics [KMO10, JNS13, Har14, HW14, NNS$^+$14, JN14, NJS13, CLS15b, JO14, AGJ14b, YCS13, LWB13, QSW14, LWB13, AAJ$^+$13, AAN13, AGM13, AGMM15, ABGM14], the initialization plus local refinement strategy mostly differs from practice, where special initializations are not needed or unavailable. The analytic techniques developed there are mostly fragmented and highly specialized. The analytic and algorithmic framework we developed here holds promise for providing a coherent account of these problems. It is interesting to see to what extent we can streamline and generalize the framework.

Our motivating experiment on real images in Chapter 1 remains mysterious. If we were to believe that real image data are "nice" and our objective there does not have spurious local minimizers either, it is surprising ADM would escape all other critical points – this is not predicted by classic or even modern theories. One reasonable place to start is to look at how gradient descent algorithms with generic initializations can escape local maximizers and saddle points (at least with high probability). The recent work [GHJY15] has showed that randomly perturbing each iterate can help gradient descent method to escape from saddle points with high probability. It is interesting to know whether similar results can be obtained for deterministic gradient descent algorithms with random initialization; see, e.g., [LSJR16] for such an attempt. The continuous counterpart seems well understood; see, e.g., [HMG94] for discussions of Morse-Bott theorem and gradient flow convergence.

[1] The recent work [AGMM15] on overcomplete DR has used a similar idea. The marginalization taken there is near to the global optimum of one variable, where the function is well-behaved. Studying the global properties of the marginalization may introduce additional technical challenges.

Chapter 9

Proof of the Function Landscape

Another roof, another proof.

Paul Erdős

In this chapter, we provide complete proofs for technical results stated in Chapter 4. Before that, let us introduce some notations and common results that will be used later throughout this chapter. Since we deal with BG random variables and random vectors, it is often convenient to write such vector explicitly as $\boldsymbol{x} = [\Omega_1 v_1, \dots, \Omega_n v_n] = \boldsymbol{\Omega} \odot \boldsymbol{v}$, where $\Omega_1, \dots, \Omega_n$ are i.i.d. Bernoulli random variables and $v_1, \dots, v_n$ are i.i.d. standard normal. For a particular realization of such random vector, we will denote the support as $\mathcal{I} \subset [n]$. Due to the particular coordinate map in use, we will often refer to subset $\mathcal{J} \doteq \mathcal{I} \setminus \{n\}$ and the random vectors $\overline{\boldsymbol{x}} \doteq [\Omega_1 v_1, \dots, \Omega_{n-1} v_{n-1}]$ and $\overline{\boldsymbol{v}} \doteq [v_1, \dots, v_{n-1}]$ in $\mathbb{R}^{n-1}$. By Lemma B.1, it is not hard to see that

$$\nabla_{\boldsymbol{w}} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right) = \tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\left(\overline{\boldsymbol{x}} - \frac{x_n}{q_n\left(\boldsymbol{w}\right)}\boldsymbol{w}\right), \tag{9.0.1}$$

$$\begin{aligned}\nabla^2_{\boldsymbol{w}} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right) = &\frac{1}{\mu}\left[1 - \tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right]\left(\overline{\boldsymbol{x}} - \frac{x_n}{q_n\left(\boldsymbol{w}\right)}\boldsymbol{w}\right)\left(\overline{\boldsymbol{x}} - \frac{x_n}{q_n\left(\boldsymbol{w}\right)}\boldsymbol{w}\right)^* \\ &- x_n \tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\left(\frac{1}{q_n\left(\boldsymbol{w}\right)}\boldsymbol{I} + \frac{1}{q_n^3\left(\boldsymbol{w}\right)}\boldsymbol{w}\boldsymbol{w}^*\right).\end{aligned} \tag{9.0.2}$$

9.1 Proofs for Section 4.2

9.1.1 Proof of Proposition 4.5

The proof involves some delicate analysis, particularly polynomial approximation of the function $f\left(t\right) = \frac{1}{(1+t)^2}$ over $t \in [0, 1]$. This is naturally induced by the $1 - \tanh^2\left(\cdot\right)$ function. The next lemma characterizes one such approximation.

Lemma 9.1 *Consider $f(t) = \frac{1}{(1+t)^2}$ for $t \in [0,1]$. For every $T > 1$, there is a sequence $b_0, b_1, \ldots,$ with $\|\boldsymbol{b}\|_{\ell^1} = T < \infty$, such that the polynomial $p(t) = \sum_{k=0}^{\infty} b_k t^k$ satisfies*

$$\|f - p\|_{L^1[0,1]} \le \frac{1}{2\sqrt{T}}, \quad \|f - p\|_{L^\infty[0,1]} \le \frac{1}{\sqrt{T}},$$

In particular, one can choose $b_k = (-1)^k (k+1)\beta^k$ with $\beta = 1 - 1/\sqrt{T} < 1$ such that

$$p(t) = \frac{1}{(1+\beta t)^2} = \sum_{k=0}^{\infty} (-1)^k (k+1) \beta^k t^k.$$

Moreover, such sequence satisfies $0 < \sum_{k=0}^{\infty} \frac{b_k}{(1+k)^3} < \sum_{k=0}^{\infty} \frac{|b_k|}{(1+k)^3} < 2$.

Lemma 9.2 *Let $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$ and $Y \sim \mathcal{N}\left(0, \sigma_Y^2\right)$. We have*

$$\mathbb{E}\left[\left(1 - \tanh^2\left(\frac{X+Y}{\mu}\right)\right) X^2 \mathbb{1}_{X+Y>0}\right] \le$$
$$\frac{1}{\sqrt{2\pi}} \frac{\mu \sigma_X^2 \sigma_Y^2}{\left(\sigma_X^2 + \sigma_Y^2\right)^{3/2}} + \frac{\mu^3 \sigma_X^2 \sigma_Y^2}{\left(\sigma_X^2 + \sigma_Y^2\right)^{3/2}} + \frac{3}{4\sqrt{2\pi}} \frac{\sigma_X^2 \mu^3}{\left(\sigma_X^2 + \sigma_Y^2\right)^{5/2}} \left(3\mu^2 + 4\sigma_X^2\right).$$

Proof For $x + y > 0$, let $z = \exp\left(-2\frac{x+y}{\mu}\right) \in [0,1]$, then $1 - \tanh^2\left(\frac{x+y}{\mu}\right) = \frac{4z}{(1+z)^2}$. Fix any $T > 1$ to be determined later, by Lemma 9.1, we choose the polynomial $p_\beta(z) = \frac{1}{(1+\beta z)^2}$ with $\beta = 1 - 1/\sqrt{T}$ to upper bound $f(z) = \frac{1}{(1+z)^2}$. So we have

$$\begin{aligned}
\mathbb{E}\left[\left(1 - \tanh^2\left(\frac{X+Y}{\mu}\right)\right) X^2 \mathbb{1}_{X+Y>0}\right] &= 4\mathbb{E}\left[Z f(Z) X^2 \mathbb{1}_{X+Y>0}\right] \\
&\le 4\mathbb{E}\left[Z p_\beta(Z) X^2 \mathbb{1}_{X+Y>0}\right] \\
&= 4 \sum_{k=0}^{\infty} \left\{ b_k \mathbb{E}\left[Z^{k+1} X^2 \mathbb{1}_{X+Y>0}\right] \right\},
\end{aligned}$$

where $b_k = (-1)^k (k+1) \beta^k$, and the exchange of infinite summation and expectation above is justified in view that

$$\sum_{k=0}^{\infty} |b_k| \, \mathbb{E}\left[Z^{k+1} X^2 \mathbb{1}_{X+Y>0}\right] \le \sum_{k=0}^{\infty} |b_k| \, \mathbb{E}\left[X^2 \mathbb{1}_{X+Y>0}\right] \le \sigma_X^2 \sum_{k=0}^{\infty} |b_k| < \infty$$

and the dominated convergence theorem (see, e.g., theorem 2.24 and 2.25 of [Fol99]). By Lemma B.10, we have

$$\sum_{k=0}^{\infty} \left\{ b_k \mathbb{E}\left[Z^{k+1} X^2 \mathbb{1}_{X+Y>0}\right] \right\}$$

$$\begin{aligned}
&= \sum_{k=0}^{\infty} (-\beta)^k (k+1) \left[\left(\sigma_X^2 + \frac{4(k+1)^2}{\mu^2}\sigma_X^4 \right) \exp\left(\frac{2(k+1)^2}{\mu^2} \left(\sigma_X^2+\sigma_Y^2\right) \right) \Phi^c \left(\frac{2(k+1)}{\mu}\sqrt{\sigma_X^2+\sigma_Y^2}\right) \right. \\
&\qquad \left. - \frac{2(k+1)}{\mu} \frac{\sigma_X^4}{\sqrt{2\pi}\sqrt{\sigma_X^2+\sigma_Y^2}} \right] \\
&\le \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} (-\beta)^k (k+1) \left[\frac{\sigma_X^2 \mu}{2(k+1)\sqrt{\sigma_X^2+\sigma_Y^2}} - \frac{\sigma_X^2\mu^3}{8(k+1)^3\left(\sigma_X^2+\sigma_Y^2\right)^{3/2}} - \frac{\mu\sigma_X^4}{2(k+1)\left(\sigma_X^2+\sigma_Y^2\right)^{3/2}} \right] \\
&\qquad + \frac{3}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \beta^k (k+1) \left(\sigma_X^2 + \frac{4(k+1)^2}{\mu^2}\sigma_X^4 \right) \frac{\mu^5}{32(k+1)^5\left(\sigma_X^2+\sigma_Y^2\right)^{5/2}},
\end{aligned}$$

where we have applied Type I upper and lower bounds for $\Phi^c(\cdot)$ to even k and odd k respectively and rearrange the terms to obtain the last line. Using the following estimates (see Lemma 9.1)

$$\sum_{k=0}^{\infty} (-\beta)^k = \frac{1}{1+\beta}, \quad \sum_{k=0}^{\infty} \frac{b_k}{(k+1)^3} \ge 0, \quad \sum_{k=0}^{\infty} \frac{|b_k|}{(k+1)^5} \le \sum_{k=0}^{\infty} \frac{|b_k|}{(k+1)^3} \le 2,$$

we obtain

$$\sum_{k=0}^{\infty} \left\{ b_k \mathbb{E}\left[Z^{k+1} X^2 \mathbb{1}_{X+Y>0} \right] \right\} \le \frac{1}{2\sqrt{2\pi}} \frac{\mu\sigma_X^2\sigma_Y^2}{\left(\sigma_X^2+\sigma_Y^2\right)^{3/2}} \frac{1}{1+\beta} + \frac{3}{16\sqrt{2\pi}} \frac{\sigma_X^2\mu^3}{\left(\sigma_X^2+\sigma_Y^2\right)^{5/2}} \left(3\mu^2 + 4\sigma_X^2\right).$$

Noticing $\frac{1}{1+\beta} < \frac{1}{2} + \frac{1}{2\sqrt{T}}$ and choosing $T = \mu^{-4}$, we obtain the desired result. ■

Lemma 9.3 *Let $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$ and $Y \sim \mathcal{N}\left(0, \sigma_Y^2\right)$. We have*

$$\begin{aligned}
&\mathbb{E}\left[\tanh\left(\frac{X+Y}{\mu} \right) X \right] \ge \\
&\qquad \frac{2\sigma_X^2}{\sqrt{2\pi}\sqrt{\sigma_X^2+\sigma_Y^2}} - \frac{4\mu^2\sigma_X^2}{\sqrt{2\pi}\sqrt{\sigma_X^2+\sigma_Y^2}} - \frac{2\sigma_X^2\mu^2}{\sqrt{2\pi}\left(\sigma_X^2+\sigma_Y^2\right)^{3/2}} - \frac{3\sigma_X^2\mu^4}{2\sqrt{2\pi}\left(\sigma_X^2+\sigma_Y^2\right)^{5/2}}.
\end{aligned}$$

Proof By Lemma B.10, we know

$$\mathbb{E}\left[\tanh\left(\frac{X+Y}{\mu} \right) X \right] = \frac{\sigma_X^2}{\mu} \mathbb{E}\left[1 - \tanh^2\left(\frac{X+Y}{\mu} \right) \right]$$

Similar to the proof of the above lemma, for $x + y > 0$, let $z = \exp\left(-2\frac{x+y}{\mu}\right)$ and $f(z) = \frac{1}{(1+z)^2}$. Fixing any $T > 1$, we will use $4zp_\beta(z) = \frac{4z}{(1+\beta z)^2}$ to approximate the $1 - \tanh^2\left(\frac{x+y}{\mu}\right) = 4zf(z)$ function from above, where again $\beta = 1 - 1/\sqrt{T}$. So we obtain

$$\begin{aligned}
\mathbb{E}\left[1 - \tanh^2\left(\frac{X+Y}{\mu} \right) \right] &= 8\mathbb{E}\left[f(Z) Z \mathbb{1}_{X+Y>0} \right] \\
&= 8\mathbb{E}\left[p_\beta(Z) Z \mathbb{1}_{X+Y>0} \right] - 8\mathbb{E}\left[\left(p_\beta(Z) - f(Z)\right) Z \mathbb{1}_{X+Y>0} \right].
\end{aligned}$$

Now for the first term, we have

$$\mathbb{E}\left[p_\beta\left(Z\right) Z \mathbb{1}_{X+Y>0}\right] = \sum_{k=0}^{\infty} b_k \mathbb{E}\left[Z^{k+1} \mathbb{1}_{X+Y>0}\right],$$

justified as $\sum_{k=0}^{\infty} |b_k| \mathbb{E}\left[Z^{k+1}\mathbb{1}_{X+Y>0}\right] \le \sum_{k=0}^{\infty} |b_k| < \infty$ making the dominated convergence theorem (see, e.g., theorem 2.24 and 2.25 of [Fol99]) applicable. To proceed, from Lemma B.10, we obtain

$$\begin{aligned}
&\sum_{k=0}^{\infty} b_k \mathbb{E}\left[Z^{k+1}\mathbb{1}_{X+Y>0}\right] \\
&= \sum_{k=0}^{\infty} \left(-\beta\right)^k \left(k+1\right) \exp\left(\frac{2}{\mu^2}\left(k+1\right)^2\left(\sigma_X^2+\sigma_Y^2\right)\right) \Phi^c\left(\frac{2}{\mu}\left(k+1\right)\sqrt{\sigma_X^2+\sigma_Y^2}\right) \\
&\ge \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \left(-\beta\right)^k \left(k+1\right)\left(\frac{\mu}{2\left(k+1\right)\sqrt{\sigma_X^2+\sigma_Y^2}} - \frac{\mu^3}{8\left(k+1\right)^3\left(\sigma_X^2+\sigma_Y^2\right)^{3/2}}\right) \\
&\qquad - \frac{3}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \beta^k \left(k+1\right) \frac{\mu^5}{32\left(k+1\right)^5\left(\sigma_X^2+\sigma_Y^2\right)^{5/2}},
\end{aligned}$$

where we have applied Type I upper and lower bounds for $\Phi^c\left(\cdot\right)$ to odd k and even k respectively and rearrange the terms to obtain the last line. Using the following estimates (see Lemma 9.1)

$$\sum_{k=0}^{\infty} \left(-\beta\right)^k = \frac{1}{1+\beta}, \quad 0 \le \sum_{k=0}^{\infty} \frac{b_k}{\left(k+1\right)^3} \le \sum_{k=0}^{\infty} \frac{|b_k|}{\left(k+1\right)^5} \le \sum_{k=0}^{\infty} \frac{|b_k|}{\left(k+1\right)^3} \le 2,$$

we obtain

$$\sum_{k=0}^{\infty} b_k \mathbb{E}\left[Z^{k+1}\mathbb{1}_{X+Y>0}\right] \ge \frac{\mu}{2\sqrt{2\pi}\sqrt{\sigma_X^2+\sigma_Y^2}} \frac{1}{1+\beta} - \frac{\mu^3}{4\sqrt{2\pi}\left(\sigma_X^2+\sigma_Y^2\right)^{3/2}} - \frac{3\mu^5}{16\sqrt{2\pi}\left(\sigma_X^2+\sigma_Y^2\right)^{5/2}}.$$

To proceed, by Lemma B.10 and Lemma 9.1, we have

$$\mathbb{E}\left[\left(p_\beta(Z) - f(Z)\right) Z \mathbb{1}_{X+Y>0}\right] \le \left\|p - f\right\|_{L^\infty[0,1]} \mathbb{E}\left[Z\mathbb{1}_{X+Y>0}\right] \le \frac{\mu}{2\sqrt{2\pi T}\sqrt{\sigma_X^2+\sigma_Y^2}},$$

where we have also used Type I upper bound for $\Phi^c\left(\cdot\right)$. Combining the above estimates, we get

$$\mathbb{E}\left[\tanh\left(\frac{X+Y}{\mu}\right) X\right] \ge \frac{4\sigma_X^2}{\sqrt{2\pi}\sqrt{\sigma_X^2+\sigma_Y^2}}\left(\frac{1}{1+\beta} - \frac{1}{\sqrt{T}}\right) - \frac{2\sigma_X^2\mu^2}{\sqrt{2\pi}\left(\sigma_X^2+\sigma_Y^2\right)^{3/2}} - \frac{3\sigma_X^2\mu^4}{2\sqrt{2\pi}\left(\sigma_X^2+\sigma_Y^2\right)^{5/2}}.$$

Noticing $\frac{1}{1+\beta} > \frac{1}{2}$ and taking $T = \mu^{-4}$, we obtain the claimed result. ■

So we are ready to present the proof.

Proof [of Proposition 4.5] For any $i \in [n-1]$, we have

$$\int_0^1 \int_{\boldsymbol{x}} \left|\frac{\partial}{\partial w_i} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right| \mu\left(d\boldsymbol{x}\right)\ dw_i \le \int_0^1 \int_{\boldsymbol{x}} \left(|x_i| + |x_n| \frac{1}{q_n\left(\boldsymbol{w}\right)}\right) \mu\left(d\boldsymbol{x}\right)\ dw_i < \infty.$$

Hence by Lemma B.4 we obtain $\frac{\partial}{\partial w_i}\mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] = \mathbb{E}\left[\frac{\partial}{\partial w_i}h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right]$. Moreover for any $j \in [n-1]$,

$$\int_0^1 \int_{\boldsymbol{x}} \left|\frac{\partial^2}{\partial w_j \partial w_i} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right| \mu\left(d\boldsymbol{x}\right)\ dw_j \le \\ \int_0^1 \int_{\boldsymbol{x}} \left[\frac{1}{\mu}\left(|x_i| + \frac{|x_n|}{q_n\left(\boldsymbol{w}\right)}\right)\left(|x_j| + \frac{|x_n|}{q_n\left(\boldsymbol{w}\right)}\right) + |x_n|\left(\frac{1}{q_n\left(\boldsymbol{w}\right)} + \frac{1}{q_n^3\left(\boldsymbol{w}\right)}\right)\right] \mu\left(d\boldsymbol{x}\right)\ dw_i < \infty.$$

Invoking Lemma B.4 again we obtain

$$\frac{\partial^2}{\partial w_j \partial w_i}\mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] = \frac{\partial}{\partial w_j}\mathbb{E}\left[\frac{\partial}{\partial w_i}h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] = \mathbb{E}\left[\frac{\partial^2}{\partial w_j \partial w_i}h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right].$$

The above holds for any pair of $i, j \in [n-1]$, so it follows that

$$\nabla^2_{\boldsymbol{w}}\mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] = \mathbb{E}\left[\nabla^2_{\boldsymbol{w}}h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right].$$

Hence it is easy to see that

$$\begin{aligned}
&\boldsymbol{w}^*\nabla^2_{\boldsymbol{w}}\mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right]\boldsymbol{w} \\
=&\ \frac{1}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}} - \frac{x_n}{q_n\left(\boldsymbol{w}\right)}\|\boldsymbol{w}\|^2\right)^2\right] - \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\frac{x_n}{q_n^3\left(\boldsymbol{w}\right)}\|\boldsymbol{w}\|^2\right].
\end{aligned}$$

Now the first term is

$$\begin{aligned}
&\frac{1}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}} - \frac{x_n}{q_n\left(\boldsymbol{w}\right)}\|\boldsymbol{w}\|^2\right)^2\right] \\
=&\ \frac{2\left(1-\theta\right)}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}}{\mu}\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}}\right)^2\mathbb{1}_{\boldsymbol{w}^*\overline{\boldsymbol{x}}>0}\right] \\
&- \frac{4\theta}{\mu}\frac{\|\boldsymbol{w}\|^2}{q_n^2\left(\boldsymbol{w}\right)}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}} + q_n\left(\boldsymbol{w}\right)x_n}{\mu}\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}}\right)\left(q_n\left(\boldsymbol{w}\right)x_n\right)\mathbb{1}_{\boldsymbol{w}^*\overline{\boldsymbol{x}}+q_n(\boldsymbol{w})x_n>0}\right] \\
&+ \frac{2\theta}{\mu}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{w}^*_{\mathcal{J}}\overline{\boldsymbol{v}} + q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right)\right)\left(\boldsymbol{w}^*_{\mathcal{J}}\overline{\boldsymbol{v}}\right)^2\mathbb{1}_{\boldsymbol{w}^*_{\mathcal{J}}\overline{\boldsymbol{v}}+q_n(\boldsymbol{w})v_n>0}\right] \\
&+ \frac{2\theta}{\mu}\frac{\|\boldsymbol{w}\|^4}{q_n^4\left(\boldsymbol{w}\right)}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{w}^*_{\mathcal{J}}\overline{\boldsymbol{v}} + q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right)\right)\left(q_n\left(\boldsymbol{w}\right)v_n\right)^2\mathbb{1}_{\boldsymbol{w}^*_{\mathcal{J}}\overline{\boldsymbol{v}}+q_n(\boldsymbol{w})v_n>0}\right] \\
\le&\ \frac{8\left(1-\theta\right)}{\mu}\mathbb{E}\left[\exp\left(-2\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}}{\mu}\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}}\right)^2\mathbb{1}_{\boldsymbol{w}^*\overline{\boldsymbol{x}}>0}\right] \\
&+ \frac{8\theta}{\mu}\frac{\|\boldsymbol{w}\|^2}{q_n^2\left(\boldsymbol{w}\right)}\mathbb{E}\left[\exp\left(-\frac{2}{\mu}\left(\boldsymbol{w}^*\overline{\boldsymbol{x}} + q_n\left(\boldsymbol{w}\right)x_n\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}} + q_n\left(\boldsymbol{w}\right)x_n\right)^2\mathbb{1}_{\boldsymbol{w}^*\overline{\boldsymbol{x}}+q_n(\boldsymbol{w})x_n>0}\right] \\
&+ \frac{2\theta}{\mu}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{X,Y}\left[\left(1-\tanh^2\left(\frac{X+Y}{\mu}\right)\right)Y^2\mathbb{1}_{X+Y>0}\right] \\
&+ \frac{2\theta}{\mu}\frac{\|\boldsymbol{w}\|^4}{q_n^4\left(\boldsymbol{w}\right)}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{X,Y}\left[\left(1-\tanh^2\left(\frac{X+Y}{\mu}\right)\right)X^2\mathbb{1}_{X+Y>0}\right],
\end{aligned}$$

where conditioned on each support set $\mathcal{J}$, we let $X \doteq q_n\left(\boldsymbol{w}\right)v_n \sim \mathcal{N}\left(0, q_n^2\left(\boldsymbol{w}\right)\right)$ and $Y \doteq \boldsymbol{w}^*_{\mathcal{J}}\overline{\boldsymbol{v}} \sim$

$\mathcal{N}\left(0, \left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2\right)$. Noticing the fact $t \mapsto \exp\left(-2t/\mu\right)t^2$ for $t > 0$ is maximized at $t = \mu$ with maximum value $\exp\left(-2\right)\mu^2$, and in view of the estimate in Lemma 9.2, we obtain

$$
\begin{aligned}
&\frac{1}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}}-\frac{x_n}{q_n\left(\boldsymbol{w}\right)}\left\|\boldsymbol{w}\right\|^2\right)^2\right] \\
&\le 8\exp\left(-2\right)\left(1-\theta+\frac{\left\|\boldsymbol{w}\right\|^2}{q_n^2\left(\boldsymbol{w}\right)}\theta\right)\mu \\
&\quad+\frac{2\theta}{\mu}\mathbb{E}_{\mathcal{J}}\left[\frac{1}{\sqrt{2\pi}}\frac{\mu\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2 q_n^2\left(\boldsymbol{w}\right)}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3}+\frac{\mu^3\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2 q_n^2\left(\boldsymbol{w}\right)}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3}+\frac{3}{4\sqrt{2\pi}}\frac{\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2\mu^3}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^5}\left(3\mu^2+4\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2\right)\right] \\
&\quad+\frac{2\theta}{\mu}\frac{\left\|\boldsymbol{w}\right\|^4}{q_n^4\left(\boldsymbol{w}\right)}\mathbb{E}_{\mathcal{J}}\left[\frac{1}{\sqrt{2\pi}}\frac{\mu\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2 q_n^2\left(\boldsymbol{w}\right)}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3}+\frac{\mu^3\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2 q_n^2\left(\boldsymbol{w}\right)}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3}+\frac{3}{4\sqrt{2\pi}}\frac{q_n^2\left(\boldsymbol{w}\right)\mu^3}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^5}\left(3\mu^2+4q_n^2\left(\boldsymbol{w}\right)\right)\right] \\
&\le \frac{2\theta}{\sqrt{2\pi}q_n^2\left(\boldsymbol{w}\right)}\mathbb{E}_{\mathcal{J}}\left[\frac{\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3}\right]+\frac{11}{20}\mu\left(2+\frac{1}{q_n^2\left(\boldsymbol{w}\right)}\right)+2\theta\mu^2\left(1+\frac{3}{\sqrt{2\pi}q_n\left(\boldsymbol{w}\right)}+\frac{1}{q_n^3\left(\boldsymbol{w}\right)}+\frac{3}{\sqrt{2\pi}q_n^5\left(\boldsymbol{w}\right)}\right),
\end{aligned}
$$

where we have used $\mu < q_n\left(\boldsymbol{w}\right) \le \left\|\boldsymbol{q}_{\mathcal{I}}\right\|$ and $\left\|\boldsymbol{w}_{\mathcal{J}}\right\| \le \left\|\boldsymbol{q}_{\mathcal{I}}\right\|$ and $\left\|\boldsymbol{w}\right\| \le 1$ and $\theta \in \left(0, 1/2\right)$ to simplify the intermediate quantities to obtain the last line. Similarly for the second term, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\frac{x_n}{q_n^3\left(\boldsymbol{w}\right)}\left\|\boldsymbol{w}\right\|^2\right] \\
&=\frac{\left\|\boldsymbol{w}\right\|^2\theta}{q_n^4\left(\boldsymbol{w}\right)}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\tanh\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right)x_n q_n\left(\boldsymbol{w}\right)\right] \\
&\ge\frac{\left\|\boldsymbol{w}\right\|^2\theta}{q_n^4\left(\boldsymbol{w}\right)}\mathbb{E}_{\mathcal{J}}\left[\frac{2q_n^2\left(\boldsymbol{w}\right)}{\sqrt{2\pi}\left\|\boldsymbol{q}_{\mathcal{I}}\right\|}-\frac{4\mu^2 q_n^2\left(\boldsymbol{w}\right)}{\sqrt{2\pi}\left\|\boldsymbol{q}_{\mathcal{I}}\right\|}-\frac{2q_n^2\left(\boldsymbol{w}\right)\mu^2}{\sqrt{2\pi}\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3}-\frac{3q_n^2\left(\boldsymbol{w}\right)\mu^4}{2\sqrt{2\pi}\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^5}\right] \\
&\ge\sqrt{\frac{2}{\pi}}\frac{\theta}{q_n^2\left(\boldsymbol{w}\right)}\mathbb{E}_{\mathcal{J}}\left[\frac{\left\|\boldsymbol{w}\right\|^2}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|}\right]-\frac{4\theta\mu^2}{\sqrt{2\pi}}\left(\frac{1}{q_n^3\left(\boldsymbol{w}\right)}+\frac{1}{q_n^5\left(\boldsymbol{w}\right)}\right).
\end{aligned}
$$

Collecting the above estimates, we obtain

$$
\begin{aligned}
&\boldsymbol{w}^*\nabla_{\boldsymbol{w}}^2\mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right]\boldsymbol{w} \\
&\le\sqrt{\frac{2}{\pi}}\frac{\theta}{q_n^2\left(\boldsymbol{w}\right)}\mathbb{E}_{\mathcal{J}}\left[\frac{\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3}-\frac{\left\|\boldsymbol{w}\right\|^2\left(\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2+q_n^2\left(\boldsymbol{w}\right)\right)}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3}\right] \\
&\qquad+\frac{11}{20}\mu\left(2+\frac{1}{q_n^2\left(\boldsymbol{w}\right)}\right)+2\theta\mu^2\left(1+\frac{3}{\sqrt{2\pi}q_n\left(\boldsymbol{w}\right)}+\frac{2}{q_n^3\left(\boldsymbol{w}\right)}+\frac{5}{\sqrt{2\pi}q_n^5\left(\boldsymbol{w}\right)}\right) \\
&\le-\sqrt{\frac{2}{\pi}}\theta\mathbb{E}\left[\frac{\left\|\boldsymbol{w}_{\mathcal{J}^c}\right\|^2}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3}\right]+\frac{11}{10}\mu+\frac{11}{20}\frac{\mu}{q_n\left(\boldsymbol{w}\right)}+2\theta\mu^2\left(1+\frac{6}{q_n^5\left(\boldsymbol{w}\right)}\right) \\
&\le-\sqrt{\frac{2}{\pi}}\theta\left(1-\theta\right)\left\|\boldsymbol{w}\right\|^2\mathbb{E}\left[\frac{1}{\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3}\right]+\frac{11}{10}\mu+\frac{11}{20}\frac{\mu}{q_n^2\left(\boldsymbol{w}\right)}+2\theta\mu^2\left(1+\frac{6}{q_n^5\left(\boldsymbol{w}\right)}\right),
\end{aligned}
\tag{9.1.1}
$$

where to obtain the last line we have invoked the association inequality in Lemma B.3, as both $\left\|\boldsymbol{w}_{\mathcal{J}^c}\right\|^2$ and

$1/\left\|\boldsymbol{q}_{\mathcal{I}}\right\|^3$ both coordinatewise nonincreasing w.r.t. the index set. Substituting the upper bound for μ into (9.1.1) and noting $R_h \le \|\boldsymbol{w}\|$ and also noting the fact $q_n\left(\boldsymbol{w}\right) \ge \frac{1}{2\sqrt{n}}$ (implied by the assumption $\|\boldsymbol{w}\| \le \sqrt{\frac{4n-1}{4n}}$), we obtain the claimed result. ■

9.1.2 Proof of Proposition 4.6

Proof By similar consideration as the above proof, the following is justified:

$$\nabla_{\boldsymbol{w}} \mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] = \mathbb{E}\left[\nabla_{\boldsymbol{w}} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right].$$

Now consider

$$\begin{aligned}\boldsymbol{w}^*\nabla\mathbb{E}\left[h_\mu(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x})\right] &= \nabla\mathbb{E}\left[\boldsymbol{w}^* h_\mu(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x})\right] \\ &= \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}}\right)\right] - \frac{\|\boldsymbol{w}\|^2}{q_n}\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)x_n\right]. \end{aligned} \tag{9.1.2}$$

For (9.1.2), we next provide a lower bound for the first expectation and an upper bound for the second expectation. For the first, we have

$$\begin{aligned}&\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}}\right)\right] \\ &= \theta\mathbb{E}_{\mathcal{J}}\left[\mathbb{E}_{\boldsymbol{v}}\left[\tanh\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}} + q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right)\left(\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}\right)\right]\right] + (1-\theta)\mathbb{E}_{\mathcal{J}}\left[\mathbb{E}_{\boldsymbol{v}}\left[\tanh\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}}{\mu}\right)\left(\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}\right)\right]\right] \\ &= \theta\mathbb{E}_{\mathcal{J}}\left[\mathbb{E}_{X,Y}\left[\tanh\left(\frac{X+Y}{\mu}\right)Y\right]\right] \ + \ (1-\theta)\mathbb{E}_{\mathcal{J}}\left[\mathbb{E}_{Y}\left[\tanh\left(\frac{Y}{\mu}\right)Y\right]\right],\end{aligned}$$

where $X \doteq q_n\left(\boldsymbol{w}\right)v_n \sim \mathcal{N}\left(0, q_n^2\left(\boldsymbol{w}\right)\right)$ and $Y \doteq \boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}} \sim \mathcal{N}\left(0, \left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2\right)$. Now by Lemma B.3 we obtain

$$\mathbb{E}\left[\tanh\left(\frac{X+Y}{\mu}\right)Y\right] \ge \mathbb{E}\left[\tanh\left(\frac{X+Y}{\mu}\right)\right]\mathbb{E}\left[Y\right] = 0,$$

as $\tanh\left(\frac{X+Y}{\mu}\right)$ and X are both coordinatewise nondecreasing function of X and Y. Using the $\tanh\left(z\right) \ge \left(1-\exp\left(-2z\right)\right)/2$ lower bound for $z > 0$ and integral results in Lemma B.10, we obtain

$$\begin{aligned}\mathbb{E}\left[\tanh\left(\frac{Y}{\mu}\right)Y\right] &= 2\mathbb{E}\left[\tanh\left(\frac{Y}{\mu}\right)Y\mathbb{1}_{Y>0}\right] \\ &\ge \mathbb{E}\left[\left(1-\exp\left(-\frac{2Y}{\mu}\right)\right)Y\mathbb{1}_{Y>0}\right] \\ &= \frac{2\sigma_Y^2}{\mu}\exp\left(\frac{2\sigma_Y^2}{\mu^2}\right)\Phi^c\left(\frac{2\sigma_Y}{\mu}\right)\end{aligned}$$

$$\geq \frac{2\sigma_Y^2}{\mu\sqrt{2\pi}}\left(\sqrt{1+\frac{\sigma_Y^2}{\mu^2}}-\frac{\sigma_Y}{\mu}\right)$$
$$\geq \frac{2\sigma_Y^2}{\mu\sqrt{2\pi}}\left(\sqrt{1+\frac{\left\|\boldsymbol{w}\right\|^2}{\mu^2}}-\frac{\left\|\boldsymbol{w}\right\|}{\mu}\right),$$

where at the second last inequality we have used Type III lower bound for Gaussian upper tail $\Phi^c(\cdot)$ (Lemma B.5), and at the last we have used the fact that $t \mapsto \sqrt{1+t^2}-t$ is a monotonic decreasing function over $t > 0$ and that $\sigma_Y = \left\|\boldsymbol{w}_{\mathcal{J}}\right\| \leq \left\|\boldsymbol{w}\right\|$. Collecting the above estimates, we have

$$\begin{aligned}\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu}\right)(\boldsymbol{w}^*\overline{\boldsymbol{x}})\right] &\geq (1-\theta)\,\mathbb{E}_{\mathcal{J}}\left[\frac{2\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2}{\mu\sqrt{2\pi}}\left(\sqrt{1+\frac{\left\|\boldsymbol{w}\right\|_2^2}{\mu^2}}-\frac{\left\|\boldsymbol{w}\right\|}{\mu}\right)\right]\\ &\geq (1-\theta)\,\mathbb{E}_{\mathcal{J}}\left[\frac{2\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2}{\mu\sqrt{2\pi}}\frac{\mu}{10\left\|\boldsymbol{w}\right\|}\right]\\ &\geq \frac{\theta(1-\theta)\left\|\boldsymbol{w}\right\|}{5\sqrt{2\pi}}, \end{aligned} \tag{9.1.3}$$

where at the second line we have used the assumption that $\left\|\boldsymbol{w}\right\| \geq \frac{\mu}{6\sqrt{2}}$ and also the fact that $\sqrt{1+x^2} \geq x+\frac{1}{10x}$ for $x \geq \frac{1}{6\sqrt{2}}$.

For the second expectation of (9.1.2), we have

$$\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu}\right)x_n\right] \leq \theta\mathbb{E}\left[\left|\tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu}\right)\right|\left|v_n\right|\right] \leq \theta\sqrt{\frac{2}{\pi}}, \tag{9.1.4}$$

as $\tanh(\cdot)$ is bounded by one in magnitude. Plugging the results of (9.1.3) and (9.1.4) into (9.1.2) and noticing that $q_n(\boldsymbol{w})^2+\left\|\boldsymbol{w}\right\|^2 = 1$ we obtain

$$\boldsymbol{w}^*\nabla\mathbb{E}\left[h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x})\right] \;\geq\; \frac{\theta\left\|\boldsymbol{w}\right\|}{\sqrt{2\pi}}\left[\frac{1-\theta}{5}-\frac{2\left\|\boldsymbol{w}\right\|}{\sqrt{1-\left\|\boldsymbol{w}\right\|^2}}\right] \geq \frac{\theta(1-\theta)\left\|\boldsymbol{w}\right\|}{10\sqrt{2\pi}},$$

where we have invoked the assumption that $\left\|\boldsymbol{w}\right\| \leq \frac{1}{10\sqrt{5}}(1-\theta)$ to provide the upper bound $\frac{2\left\|\boldsymbol{w}\right\|}{\sqrt{1-\left\|\boldsymbol{w}\right\|^2}} \leq \frac{1}{10}(1-\theta)$. We then choose the particular ranges as stated for μ and θ to ensure $r_g < R_g$, completing the proof. ■

9.1.3 Proof of Proposition 4.7

Proof By consideration similar to the above proof, we can exchange the Hessian operator and expectation, i.e.,

$$\nabla^2_{\boldsymbol{w}} \mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] = \mathbb{E}\left[\nabla^2_{\boldsymbol{w}} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right].$$

We are interested in the expected Hessian matrix

$$\begin{aligned}\nabla^2_{\boldsymbol{w}} \mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] = {} & \frac{1}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)x}{\mu}\right)\right)\left(\overline{\boldsymbol{x}}-\frac{x_n}{q_n\left(\boldsymbol{w}\right)}\boldsymbol{w}\right)\left(\overline{\boldsymbol{x}}-\frac{x_n}{q_n\left(\boldsymbol{w}\right)}\boldsymbol{w}\right)^*\right] \\ & - \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\left(\frac{x_n}{q_n\left(\boldsymbol{w}\right)}\boldsymbol{I}+\frac{x_n}{q_n^3\left(\boldsymbol{w}\right)}\boldsymbol{w}\boldsymbol{w}^*\right)\right]\end{aligned}$$

in the region that $0 \le \|\boldsymbol{w}\| \le \frac{\mu}{4\sqrt{2}}$.

When $\boldsymbol{w} = \boldsymbol{0}$, by Lemma B.10, we have

$$\begin{aligned} & \mathbb{E}\left[\nabla^2_{\boldsymbol{w}} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right]\Big|_{\boldsymbol{w}=0} \\ = {} & \frac{1}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{x_n}{\mu}\right)\right)\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^*\right] - \mathbb{E}\left[\tanh\left(\frac{x_n}{\mu}\right)x_n\right]\boldsymbol{I} \\ = {} & \frac{\theta(1-\theta)}{\mu}\boldsymbol{I} + \frac{\theta^2}{\mu}\mathbb{E}_{v_n}\left[1-\tanh^2\left(\frac{v_n}{\mu}\right)\right]\boldsymbol{I} - \frac{\theta}{\mu}\mathbb{E}_{v_n}\left[1-\tanh^2\left(\frac{v_n}{\mu}\right)\right]\boldsymbol{I} \\ = {} & \frac{\theta(1-\theta)}{\mu}\mathbb{E}_{v_n}\left[\tanh^2\left(\frac{q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right)\right]\boldsymbol{I}.\end{aligned}$$

Simple calculation based on Lemma B.10 shows

$$\mathbb{E}_{v_n}\left[\tanh^2\left(\frac{v_n}{\mu}\right)\right] \ge 2\left(1-4\exp\left(\frac{2}{\mu^2}\right)\Phi^c\left(\frac{2}{\mu}\right)\right) \ge 2\left(1-\frac{2}{\sqrt{2\pi}}\mu\right).$$

Invoking the assumptions $\mu \le \frac{1}{20\sqrt{n}} \le 1/20$ and $\theta < 1/2$, we obtain

$$\mathbb{E}\left[\nabla^2_{\boldsymbol{w}} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right]\Big|_{\boldsymbol{w}=0} \succeq \frac{\theta\left(1-\theta\right)}{\mu}\left(2-\frac{4}{\sqrt{2\pi}}\mu\right)\boldsymbol{I} \succeq \frac{\theta}{\mu}\left(1-\frac{1}{10\sqrt{2\pi}}\right)\boldsymbol{I}.$$

When $0 < \|\boldsymbol{w}\| \le \frac{\mu}{4\sqrt{2}}$, we aim to derive a semidefinite lower bound for

$$\begin{aligned} & \mathbb{E}\left[\nabla^2_{\boldsymbol{w}} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] \\ = {} & \frac{1}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^*\right] - \frac{1}{q_n^2\left(\boldsymbol{w}\right)}\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)q_n\left(\boldsymbol{w}\right)x_n\right]\boldsymbol{I} \\ & - \frac{1}{\mu q_n^2\left(\boldsymbol{w}\right)}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)q_n\left(\boldsymbol{w}\right)x_n\left(\boldsymbol{w}\overline{\boldsymbol{x}}^*+\overline{\boldsymbol{x}}\boldsymbol{w}^*\right)\right]\end{aligned}$$

$$+\frac{1}{q_n^4\left(\boldsymbol{w}\right)}\left\{\frac{1}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)\left(q_n\left(\boldsymbol{w}\right)x_n\right)^2\right]-\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)q_n\left(\boldsymbol{w}\right)x_n\right]\right\}\boldsymbol{w}\boldsymbol{w}^*. \tag{9.1.5}$$

We will first provide bounds for the last two lines and then tackle the first which is slightly more tricky. For the second line, we have

$$\begin{aligned}
&\frac{1}{\mu q_n^2\left(\boldsymbol{w}\right)}\left\|\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)q_n\left(\boldsymbol{w}\right)x_n\left(\boldsymbol{w}\overline{\boldsymbol{x}}^*+\overline{\boldsymbol{x}}\boldsymbol{w}^*\right)\right]\right\|\\
\leq\;&\frac{2}{\mu q_n^2\left(\boldsymbol{w}\right)}\left\|\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)q_n\left(\boldsymbol{w}\right)x_n\bar{\boldsymbol{x}}\right]\boldsymbol{w}^*\right\|\\
\leq\;&\frac{2}{\mu q_n^2\left(\boldsymbol{w}\right)}\left\|\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)q_n\left(\boldsymbol{w}\right)x_n\overline{\boldsymbol{x}}\right]\right\|\left\|\boldsymbol{w}\right\|\\
\leq\;&\frac{2}{\mu q_n\left(\boldsymbol{w}\right)}\theta^2\mathbb{E}\left[\left|v_n\right|\right]\mathbb{E}\left[\left\|\overline{\boldsymbol{v}}\right\|\right]\left\|\boldsymbol{w}\right\|\\
\leq\;&\frac{4\theta^2}{\pi\mu q_n\left(\boldsymbol{w}\right)}\sqrt{n}\left\|\boldsymbol{w}\right\|\leq\frac{\theta}{\mu}\frac{4\theta\sqrt{n}\left\|\boldsymbol{w}\right\|}{\pi\sqrt{1-\left\|\boldsymbol{w}\right\|^2}}\leq\frac{\theta}{\mu}\frac{1}{40\pi},
\end{aligned}$$

where from the third to the fourth line we have used $\left\|1-\tanh^2\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}}{\mu}\right)\right\|\leq 1$, Jensen's inequality for the $\left\|\cdot\right\|$ function, and independence of x_n and $\overline{\boldsymbol{x}}$, and to obtain the last bound we have invoked the $\left\|\boldsymbol{w}\right\|\leq\frac{\mu}{4\sqrt{2}}$, $\mu\leq\frac{1}{20\sqrt{n}}$, and $\theta<\frac{1}{2}$ assumptions. For the third line in (9.1.5), by Lemma B.1 and Lemma B.10,

$$\begin{aligned}
&\left|\frac{1}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)\left(q_n\left(\boldsymbol{w}\right)x_n\right)^2\right]-\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)q_nx_n\right]\right|\\
=\;&\left|\frac{\theta}{\mu}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right)\left(q_n\left(\boldsymbol{w}\right)v_n\right)^2\right)\right]\right.\\
&\qquad\left.-\theta\mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\tanh\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right)q_n\left(\boldsymbol{w}\right)v_n\right]\right|\\
=\;&\frac{\theta}{\mu}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n\left(\boldsymbol{w}s\right)v_n}{\mu}\right)\right)\left(\left(q_n\left(\boldsymbol{w}\right)v_n\right)^2+q_n^2\left(\boldsymbol{w}\right)\right)\right]\\
\leq\;&\frac{8\theta}{\mu}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\exp\left(-\frac{2}{\mu}\left(\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n\left(\boldsymbol{w}\right)v_n\right)\right)\left(\left(q_n\left(\boldsymbol{w}\right)v_n\right)^2+q_n^2\left(\boldsymbol{w}\right)\right)\mathbb{1}_{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n(\boldsymbol{w})v_n>0}\right]\\
\leq\;&\frac{8\theta}{\sqrt{2\pi}}\mathbb{E}_{\mathcal{J}}\left[\frac{q_n^2\left(\boldsymbol{w}\right)}{\sqrt{q_n^2\left(\boldsymbol{w}\right)+\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2}}\right]\;\leq\;\frac{8\theta q_n\left(\boldsymbol{w}\right)}{\sqrt{2\pi}}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\frac{1}{q_n^4\left(\boldsymbol{w}\right)}\left\{\frac{1}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\boldsymbol{x}}{\mu}\right)\right)\left(q_nx_n\right)^2\right]-\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\boldsymbol{x}}{\mu}\right)q_nx_n\right]\right\}\boldsymbol{w}\boldsymbol{w}^*\\
\succeq\;&-\frac{8\theta}{q_n^3\left(\boldsymbol{w}\right)\sqrt{2\pi}}\left\|\boldsymbol{w}\right\|^2\boldsymbol{I}\succeq-\frac{\theta}{\mu}\left(\frac{64n^{3/2}\mu\left\|\boldsymbol{w}\right\|^2}{q_n^3\left(\boldsymbol{w}\right)\sqrt{2\pi}}\right)\boldsymbol{I}\succeq-\frac{\theta}{\mu}\frac{1}{4000\sqrt{2\pi}}\boldsymbol{I},
\end{aligned}$$

where we have again used $\left\|\boldsymbol{w}\right\|\leq\frac{\mu}{4\sqrt{2}}$, $\mu\leq\frac{1}{20\sqrt{n}}$, and $q_n\left(\boldsymbol{w}\right)\geq\frac{1}{2\sqrt{n}}$ assumptions to simplify the final bound.

To derive a lower bound for the first line of (9.1.5), we lower bound the first term and upper bound the second. The latter is easy: using Lemma B.1 and Lemma B.10,

$$\begin{aligned}
&\frac{1}{q_n^2\left(\boldsymbol{w}\right)}\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)q_n\left(\boldsymbol{w}\right)x_n\right] \\
&= \frac{\theta}{\mu}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[1-\tanh^2\left[\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right]\right] \\
&\le \frac{8\theta}{\mu}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\exp\left(-2\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right)\mathbb{1}_{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n(\boldsymbol{w})v_n>0}\right] \\
&\le \frac{4\theta}{\sqrt{2\pi}q_n\left(\boldsymbol{w}\right)} \le \frac{\theta}{\mu}\frac{8\sqrt{n}\mu}{\sqrt{2\pi}} \le \frac{\theta}{\mu}\frac{2}{5\sqrt{2\pi}},
\end{aligned}$$

where we have again used assumptions that $q_n\left(\boldsymbol{w}\right) \ge \frac{1}{2\sqrt{n}}$ and $\mu \le \frac{1}{20\sqrt{n}}$ to simplify the last bound. To lower bound the first term, first note that

$$\frac{1}{\mu}\mathbb{E}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^*\right] \succeq \frac{1-\theta}{\mu}\mathbb{E}_{\overline{\boldsymbol{x}}}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}}{\mu}\right)\right)\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^*\right].$$

We set out to lower bound the expectation as

$$\mathbb{E}_{\overline{\boldsymbol{x}}}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}}{\mu}\right)\right)\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^*\right] \succeq \theta\beta\boldsymbol{I}$$

for some scalar $\beta > 0$. Suppose $\boldsymbol{w}$ has $k \in [n-1]$ nonzeros, w.l.o.g., further assume the first k elements of $\boldsymbol{w}$ are these nonzeros. It is easy to see the expectation above has a block diagonal structure $\mathrm{diag}\left(\boldsymbol{\Sigma}; \alpha\theta\boldsymbol{I}_{n-1-k}\right)$, where

$$\alpha \doteq \mathbb{E}_{\overline{\boldsymbol{x}}}\left[\left(1-\tanh^2\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}}{\mu}\right)\right)\right].$$

So in order to derive the $\theta\beta\boldsymbol{I}$ lower bound as desired, it is sufficient to show $\boldsymbol{\Sigma} \succeq \theta\beta\boldsymbol{I}$ for some $0 < \beta < 1$, i.e., letting $\widetilde{\boldsymbol{w}} \in \mathbb{R}^k$ be the subvector of nonzero elements,

$$\mathbb{E}_{\widetilde{\boldsymbol{x}}\sim_{i.i.d.}\mathrm{BG}(\theta)}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}^*\widetilde{\boldsymbol{x}}}{\mu}\right)\right)\widetilde{\boldsymbol{x}}\,\widetilde{\boldsymbol{x}}^*\right] \succeq \theta\beta\boldsymbol{I},$$

which is equivalent to that for all $\boldsymbol{z} \in \mathbb{R}^k$ such that $\|\boldsymbol{z}\| = 1$,

$$\mathbb{E}_{\widetilde{\boldsymbol{x}}\sim_{i.i.d.}\mathrm{BG}(\theta)}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}^*\widetilde{\boldsymbol{x}}}{\mu}\right)\right)\left(\widetilde{\boldsymbol{x}}^*\boldsymbol{z}\right)^2\right] \ge \theta\beta.$$

It is then sufficient to show that for any nontrivial support set $\mathcal{S} \subset [k]$ and any vector $\boldsymbol{z} \in \mathbb{R}^k$ such that $\mathrm{supp}\left(\boldsymbol{z}\right) = \mathcal{S}$ with $\|\boldsymbol{z}\| = 1$,

$$\mathbb{E}_{\widetilde{\boldsymbol{v}}\sim_{i.i.d.}\mathcal{N}(0,1)}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}}{\mu}\right)\right)\left(\widetilde{\boldsymbol{v}}^*\boldsymbol{z}\right)^2\right] \ge \beta.$$

To see the implication, suppose the latter claimed holds, then for any $\boldsymbol{z}$ with unit norm,

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\boldsymbol{x}}\sim_{i.i.d.}\mathrm{BG}(\theta)}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}^*\widetilde{\boldsymbol{x}}}{\mu}\right)\right)\left(\widetilde{\boldsymbol{x}}^*\boldsymbol{z}\right)^2\right] \\
&= \sum_{s=1}^{k}\theta^s\left(1-\theta\right)^{k-s}\sum_{\mathcal{S}\in\binom{[k]}{s}}\mathbb{E}_{\widetilde{\boldsymbol{v}}\sim_{i.i.d.}\mathcal{N}(0,1)}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}}{\mu}\right)\right)\left(\widetilde{\boldsymbol{v}}^*\boldsymbol{z}_{\mathcal{S}}\right)^2\right] \\
&\ge \sum_{s=1}^{k}\theta^s\left(1-\theta\right)^{k-s}\sum_{\mathcal{S}\in\binom{[k]}{s}}\beta\left\|\boldsymbol{z}_{\mathcal{S}}\right\|^2 = \beta\mathbb{E}_{\mathcal{S}}\left[\left\|\boldsymbol{z}_{\boldsymbol{S}}\right\|^2\right] = \theta\beta.
\end{aligned}$$

Now for any fixed support set $\mathcal{S}\subset[k]$, $\boldsymbol{z}=\mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}}\boldsymbol{z}+\left(\boldsymbol{I}-\mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}}\right)\boldsymbol{z}$. So we have

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\boldsymbol{v}}\sim_{i.i.d.}\mathcal{N}(0,1)}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}}{\mu}\right)\right)\left(\widetilde{\boldsymbol{v}}^*\boldsymbol{z}\right)^2\right] \\
&= \mathbb{E}_{\widetilde{\boldsymbol{v}}}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}}{\mu}\right)\right)\left(\widetilde{\boldsymbol{v}}^*\mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}}\boldsymbol{z}\right)^2\right]+\mathbb{E}_{\widetilde{\boldsymbol{v}}}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}}{\mu}\right)\right)\left(\widetilde{\boldsymbol{v}}^*\left(\boldsymbol{I}-\mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}}\right)\boldsymbol{z}\right)^2\right] \\
&= \frac{\left(\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\boldsymbol{z}\right)^2}{\left\|\boldsymbol{w}_{\mathcal{S}}\right\|^4}\mathbb{E}_{\widetilde{\boldsymbol{v}}}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}}{\mu}\right)\right)\left(\widetilde{\boldsymbol{v}}^*\widetilde{\boldsymbol{w}}_{\mathcal{S}}\right)^2\right]+\mathbb{E}_{\widetilde{\boldsymbol{v}}}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}}{\mu}\right)\right)\right]\mathbb{E}_{\widetilde{\boldsymbol{v}}}\left[\left(\widetilde{\boldsymbol{v}}^*\left(\boldsymbol{I}-\mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}}\right)\boldsymbol{z}\right)^2\right] \\
&\ge 2\frac{\left(\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\boldsymbol{z}\right)^2}{\left\|\boldsymbol{w}_{\mathcal{S}}\right\|^4}\mathbb{E}_{\widetilde{\boldsymbol{v}}}\left[\exp\left(-\frac{2\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}}{\mu}\right)\left(\widetilde{\boldsymbol{v}}^*\widetilde{\boldsymbol{w}}_{\mathcal{S}}\right)^2\mathbb{1}_{\widetilde{\boldsymbol{v}}^*\widetilde{\boldsymbol{w}}_{\mathcal{S}}>0}\right]+2\mathbb{E}_{\widetilde{\boldsymbol{v}}}\left[\exp\left(-\frac{2\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}}{\mu}\right)\mathbb{1}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}>0}\right]\left\|\left(\boldsymbol{I}-\mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}}\right)\boldsymbol{z}\right\|^2.
\end{aligned}$$

Using expectation result from Lemma B.10, and applying Type III lower bound for Gaussian tails, we obtain

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\boldsymbol{v}}\sim_{i.i.d.}\mathcal{N}(0,1)}\left[\left(1-\tanh^2\left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\widetilde{\boldsymbol{v}}}{\mu}\right)\right)\left(\widetilde{\boldsymbol{v}}^*\boldsymbol{z}\right)^2\right] \\
&\ge \frac{1}{\sqrt{2\pi}}\left(\sqrt{4+\frac{4\left\|\widetilde{\boldsymbol{w}}_{\mathcal{S}}\right\|^2}{\mu^2}}-\frac{2\left\|\widetilde{\boldsymbol{w}}_{\mathcal{S}}\right\|}{\mu}\right)-\frac{4\left(\widetilde{\boldsymbol{w}}_{\mathcal{S}}^*\boldsymbol{z}\right)^2}{\mu\sqrt{2\pi}\left\|\widetilde{\boldsymbol{w}}_{\mathcal{S}}\right\|} \\
&\ge \frac{1}{\sqrt{2\pi}}\left(2-\frac{3}{4}\sqrt{2}\right),
\end{aligned}$$

where we have used Cauchy-Schwartz to obtain $\left(\widetilde{\boldsymbol{v}}^*\boldsymbol{z}\right)^2\le\left\|\widetilde{\boldsymbol{v}}^*\right\|^2$ and invoked the assumption $\|\boldsymbol{w}\|\le\frac{\mu}{4\sqrt{2}}$ to simplify the last bound. On the other hand, we similarly obtain

$$\alpha=\mathbb{E}_{\mathcal{J}}\mathbb{E}_{Z\sim\mathcal{N}(0,\|\boldsymbol{w}_{\mathcal{J}}\|^2)}[1-\tanh^2(Z/\mu)]\ge\frac{2}{\sqrt{2\pi}}\frac{\sqrt{4\|\boldsymbol{w}\|^2/\mu^2+4}-2\|\boldsymbol{w}\|/\mu}{2}\ge\frac{1}{\sqrt{2\pi}}\left(2-\frac{1}{2}\sqrt{2}\right).$$

So we can take $\beta=\frac{1}{\sqrt{2\pi}}\left(2-\frac{3}{4}\sqrt{2}\right)<1$.

Putting together the above estimates for the case $\boldsymbol{w}\neq\boldsymbol{0}$, we obtain

$$\mathbb{E}\left[\nabla_{\boldsymbol{w}}^2 h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right]\succeq\frac{\theta}{\mu\sqrt{2\pi}}\left(1-\frac{3}{8}\sqrt{2}-\frac{\sqrt{2\pi}}{40\pi}-\frac{1}{4000}-\frac{2}{5}\right)\boldsymbol{I}\succeq\frac{1}{25\sqrt{2\pi}}\frac{\theta}{\mu}\boldsymbol{I}.$$

Hence for all $\boldsymbol{w}$, we can take the $\frac{1}{25\sqrt{2\pi}}\frac{\theta}{\mu}$ as the lower bound, completing the proof. ■

9.1.4 Proof of pointwise concentration results

To avoid clutter of notations, in this subsection we write $\boldsymbol{X}$ to mean $\boldsymbol{X}_0$; similarly $\boldsymbol{x}_k$ for $(\boldsymbol{x}_0)_k$, the k-th column of $\boldsymbol{X}_0$. The function $g(\boldsymbol{w})$ means $g(\boldsymbol{w}; \boldsymbol{X}_0)$. We first establish a useful comparison lemma between random i.i.d. Bernoulli random vectors random i.i.d. normal random vectors.

Lemma 9.4 *Suppose $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^n$ are independent and obey $\boldsymbol{z} \sim_{i.i.d.} \mathrm{BG}(\theta)$ and $\boldsymbol{z}' \sim_{i.i.d.} \mathcal{N}(0,1)$. Then, for any fixed vector $\boldsymbol{v} \in \mathbb{R}^n$, it holds that*

$$\mathbb{E}\left[|\boldsymbol{v}^*\boldsymbol{z}|^m\right] \le \mathbb{E}\left[|\boldsymbol{v}^*\boldsymbol{z}'|^m\right] = \mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{v}\|^2\right)}\left[|Z|^m\right],$$
$$\mathbb{E}\left[\|\boldsymbol{z}\|^m\right] \le \mathbb{E}\left[\|\boldsymbol{z}'\|^m\right],$$

for all integers $m \ge 1$.

Proof See Section B.2.3 on Page 228. ■

Now, we are ready to prove Proposition 4.8 through Proposition 4.10 as follows.

Proof [of Proposition 4.8] Let

$$Y_k = \frac{1}{\|\boldsymbol{w}\|^2}\boldsymbol{w}^*\nabla^2 h_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)\boldsymbol{w},$$

then $\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w}}{\|\boldsymbol{w}\|^2} = \frac{1}{p}\sum_{k=1}^p Y_k$. For each Y_k ($k \in [p]$), from (9.0.2), we know that

$$Y_k = \frac{1}{\mu}\left(1-\tanh^2\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\right)\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|} - \frac{x_k(n)\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})}\right)^2 - \tanh\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\frac{x_k(n)}{q_n^3(\boldsymbol{w})}.$$

Writing $Y_k = W_k + V_k$, where

$$W_k = \frac{1}{\mu}\left(1-\tanh^2\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\right)\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|} - \frac{x_k(n)\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})}\right)^2,$$
$$V_k = -\tanh\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\frac{x_k(n)}{q_n^3(\boldsymbol{w})}.$$

Then by similar argument as in proof to Proposition 4.9, we have for all integers $m \ge 2$ that

$$\begin{aligned}\mathbb{E}\left[|W_k|^m\right] &\le \frac{1}{\mu^m}\mathbb{E}\left[\left|\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|} - \frac{x_k(n)\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})}\right|^{2m}\right] \le \frac{1}{\mu^m}\mathbb{E}_{Z\sim\mathcal{N}(0,1/q_n^2(\boldsymbol{w}))}\left[|Z|^{2m}\right] \\ &\le \frac{1}{\mu^m}(2m-1)!!(4n)^m \le \frac{m!}{2}\left(\frac{4n}{\mu}\right)^m, \\ \mathbb{E}\left[|V_k|^m\right] &\le \frac{1}{q_n^{3m}(\boldsymbol{w})}\mathbb{E}\left[|v_k(n)|^m\right] \le \left(2\sqrt{n}\right)^{3m}(m-1)!! \le \frac{m!}{2}\left(8n\sqrt{n}\right)^m,\end{aligned}$$

where we have again used the assumption that $q_n(\boldsymbol{w}) \ge \frac{1}{2\sqrt{n}}$ to simplify the result. Taking $\sigma_W^2 = 16n^2/\mu^2 \ge$

$\mathbb{E}\left[W_k^2\right]$, $R_W = 4n/\mu$ and $\sigma_V^2 = 64n^3 \geq \mathbb{E}\left[V_k^2\right]$, $R_V = 8n\sqrt{n}$, and considering $S_W = \frac{1}{p}\sum_{k=1}^p W_k$ and $S_V = \frac{1}{p}\sum_{k=1}^p V_k$, then by Lemma A.1, we obtain

$$\begin{aligned} \mathbb{P}\left[\left|S_W - \mathbb{E}\left[S_W\right]\right| \geq \frac{t}{2}\right] &\leq 2\exp\left(-\frac{p\mu^2 t^2}{128n^2 + 16n\mu t}\right), \\ \mathbb{P}\left[\left|S_V - \mathbb{E}\left[S_V\right]\right| \geq \frac{t}{2}\right] &\leq 2\exp\left(-\frac{pt^2}{512n^3 + 32n\sqrt{n}t}\right). \end{aligned}$$

Combining the above results, we obtain

$$\begin{aligned} \mathbb{P}\left[\left|\frac{1}{p}\sum_{k=1}^p X_k - \mathbb{E}\left[X_k\right]\right| \geq t\right] &= \mathbb{P}\left[\left|S_W - \mathbb{E}\left[S_W\right] + S_V - \mathbb{E}\left[S_V\right]\right| \geq t\right] \\ &\leq \mathbb{P}\left[\left|S_W - \mathbb{E}\left[S_W\right]\right| \geq \frac{t}{2}\right] + \mathbb{P}\left[\left|S_V - \mathbb{E}\left[S_V\right]\right| \geq \frac{t}{2}\right] \\ &\leq 2\exp\left(-\frac{p\mu^2 t^2}{128n^2 + 16n\mu t}\right) + 2\exp\left(-\frac{pt^2}{512n^3 + 32n\sqrt{n}t}\right) \\ &\leq 4\exp\left(-\frac{p\mu^2 t^2}{512n^2 + 32n\mu t}\right), \end{aligned}$$

provided that $\mu \leq \frac{1}{\sqrt{n}}$, as desired. ■

Proof [of Proposition 4.9] Let

$$X_k = \frac{\boldsymbol{w}^*}{\|\boldsymbol{w}\|_2}\nabla h_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right),$$

then $\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} = \frac{1}{p}\sum_{k=1}^p X_k$. For each $X_k, k \in [p]$, from (9.0.1), we know that

$$|X_k| = \left|\tanh\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|} - \frac{\|\boldsymbol{w}\|_2 x_k(n)}{q_n(\boldsymbol{w})}\right)\right| \leq \left|\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|} - \frac{\|\boldsymbol{w}\|_2 x_k(n)}{q_n(\boldsymbol{w})}\right|,$$

as the magnitude of $\tanh(\cdot)$ is bounded by one. Because $\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|_2} - \frac{\|\boldsymbol{w}\| x_k(n)}{q_n(\boldsymbol{w})} = \left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, -\frac{\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})}\right)^* \boldsymbol{x}_k$ and $\boldsymbol{x}_k \sim_{i.i.d.}$ $\mathrm{BG}(\theta)$, invoking Lemma 9.4, we obtain for every integer $m \geq 2$ that

$$\mathbb{E}\left[|X_k|^m\right] \leq \mathbb{E}_{Z\sim\mathcal{N}(0,1/q_n^2(\boldsymbol{w}))}\left[|Z|^m\right] \leq \frac{1}{q_n(\boldsymbol{w})^m}(m-1)!! \leq \frac{m!}{2}(4n)\left(2\sqrt{n}\right)^{m-2},$$

where the Gaussian moment can be looked up in Lemma B.6 and we used the fact that $(m-1)!! \leq m!/2$ and the assumption that $q_n(\boldsymbol{w}) \geq \frac{1}{2\sqrt{n}}$ to get the result. Thus, by taking $\sigma^2 = 4n \geq \mathbb{E}\left[X_k^2\right]$ and $R = 2\sqrt{n}$, and we obtain the claimed result by invoking Lemma A.1. ■

Proof [of Proposition 4.10] Let $\boldsymbol{Z}_k = \nabla_{\boldsymbol{w}}^2 h_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)$, then $\nabla_{\boldsymbol{w}}^2 g(\boldsymbol{w}) = \frac{1}{p}\sum_{k=1}^p \boldsymbol{Z}_k$. From (9.0.2), we know that

$$\boldsymbol{Z}_k = \boldsymbol{W}_k + \boldsymbol{V}_k$$

where

$$\begin{aligned}\boldsymbol{W}_k &= \frac{1}{\mu}\left(1-\tanh^2\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\right)\left(\overline{\boldsymbol{x}}_k-\frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right)\left(\overline{\boldsymbol{x}}_k-\frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right)^*\\ \boldsymbol{V}_k &= -\tanh\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\left(\frac{x_k\left(n\right)}{q_n(\boldsymbol{w})}\boldsymbol{I}+\frac{x_k\left(n\right)\boldsymbol{w}\boldsymbol{w}^*}{q_n^3(\boldsymbol{w})}\right).\end{aligned}$$

For $\boldsymbol{W}_k$, we have

$$\begin{aligned}\mathbf{0}\preceq\mathbb{E}\left[\boldsymbol{W}_k^m\right]\preceq\frac{1}{\mu^m}\mathbb{E}&\left[\left\|\overline{\boldsymbol{x}}_k-\frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right\|^{2m-2}\left(\overline{\boldsymbol{x}}_k-\frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right)\left(\overline{\boldsymbol{x}}_k-\frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right)^*\right]\\ &\preceq\frac{1}{\mu^m}\mathbb{E}\left[\left\|\overline{\boldsymbol{x}}_k-\frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right\|^{2m}\right]\boldsymbol{I}\\ &\preceq\frac{2^m}{\mu^m}\mathbb{E}\left[\left(\left\|\overline{\boldsymbol{x}}_k\right\|^2+\frac{x_k^2\left(n\right)\left\|\boldsymbol{w}\right\|^2}{q_n^2(\boldsymbol{w})}\right)^m\right]\boldsymbol{I}\\ &\preceq\frac{2^m}{\mu^m}\mathbb{E}\left[\left\|\boldsymbol{x}_k\right\|^{2m}\right]\boldsymbol{I}\ \preceq\ \frac{2^m}{\mu^m}\mathbb{E}_{Z\sim\chi^2(n)}\left[Z^m\right]\boldsymbol{I},\end{aligned}$$

where we have used the fact that $\left\|\boldsymbol{w}\right\|^2/q_n^2(\boldsymbol{w})=\left\|\boldsymbol{w}\right\|^2/(1-\left\|\boldsymbol{w}\right\|^2)\leq 1$ for $\left\|\boldsymbol{w}\right\|_2\leq\frac{1}{4}$ and Lemma 9.4 to obtain the last line. By Lemma B.7, we obtain

$$\mathbf{0}\preceq\mathbb{E}\left[\boldsymbol{W}_k^m\right]\ \preceq\ \left(\frac{2}{\mu}\right)^m\frac{m!}{2}\left(2n\right)^m\boldsymbol{I}\ =\ \frac{m!}{2}\left(\frac{4n}{\mu}\right)^m\boldsymbol{I}.$$

Taking $R_W=\frac{4n}{\mu}$ and $\boldsymbol{\sigma}_W^2=\frac{16n^2}{\mu^2}\geq\mathbb{E}\left[\boldsymbol{W}_k^2\right]$, and letting $\boldsymbol{S}_W\doteq\frac{1}{p}\sum_{k=1}^p\boldsymbol{W}_k$, by Lemma A.2, we obtain

$$\mathbb{P}\left[\left\|\boldsymbol{S}_W-\mathbb{E}\left[\boldsymbol{S}_W\right]\right\|\geq\frac{t}{2}\right]\ \leq\ 2n\exp\left(-\frac{p\mu^2t^2}{128n^2+16\mu nt}\right).$$

Similarly, for $\boldsymbol{V}_k$, we have

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{V}_k^m\right]\ &\preceq\ \left(\frac{1}{q_n(\boldsymbol{w})}+\frac{\left\|\boldsymbol{w}\right\|^2}{q_n^3(\boldsymbol{w})}\right)^m\mathbb{E}\left[\left|x_k\left(n\right)\right|^m\right]\boldsymbol{I}\\ &\preceq\ \left(8n\sqrt{n}\right)^m(m-1)!!\boldsymbol{I}\\ &\preceq\ \frac{m!}{2}\left(8n\sqrt{n}\right)^m\boldsymbol{I},\end{aligned}$$

where we have used the fact $q_n\left(\boldsymbol{w}\right)\geq\frac{1}{2\sqrt{n}}$ to simplify the result. Similar argument also shows $-\mathbb{E}\left[\boldsymbol{V}_k^m\right]\preceq m!\left(8n\sqrt{n}\right)^m\boldsymbol{I}/2$. Taking $R_V=8n\sqrt{n}$ and $\boldsymbol{\sigma}_V^2=64n^3$, and letting $\boldsymbol{S}_V\doteq\frac{1}{p}\sum_{k=1}^p\boldsymbol{V}_k$, again by Lemma A.2, we obtain

$$\mathbb{P}\left[\left\|\boldsymbol{S}_V-\mathbb{E}\left[\boldsymbol{S}_V\right]\right\|\geq\frac{t}{2}\right]\ \leq\ 2n\exp\left(-\frac{pt^2}{512n^3+32n\sqrt{n}t}\right).$$

Combining the above results, we obtain

$$\begin{aligned}\mathbb{P}\left[\left\|\frac{1}{p}\sum_{k=1}^{p}\boldsymbol{Z}_k-\mathbb{E}\left[\boldsymbol{Z}_k\right]\right\|\geq t\right] &= \mathbb{P}\left[\left\|\boldsymbol{S}_W-\mathbb{E}\left[\boldsymbol{S}_W\right]+\boldsymbol{S}_V-\mathbb{E}\left[\boldsymbol{S}_V\right]\right\|\geq t\right]\\ &\leq \mathbb{P}\left[\left\|\boldsymbol{S}_W-\mathbb{E}\left[\boldsymbol{S}_W\right]\right\|\geq\frac{t}{2}\right]+\mathbb{P}\left[\left\|\boldsymbol{S}_V-\mathbb{E}\left[\boldsymbol{S}_V\right]\right\|\geq\frac{t}{2}\right]\\ &\leq 2n\exp\left(-\frac{p\mu^2t^2}{128n^2+16\mu nt}\right)+2n\exp\left(-\frac{pt^2}{512n^3+32n\sqrt{n}t}\right)\\ &\leq 4n\exp\left(-\frac{p\mu^2t^2}{512n^2+32\mu nt}\right),\end{aligned}$$

where we have simplified the final result based on the fact that $\mu\leq\frac{1}{\sqrt{n}}$. ■

9.1.5 Proof of Lipschitz results

To avoid clutter of notations, in this subsection we write $\boldsymbol{X}$ to mean $\boldsymbol{X}_0$; similarly $\boldsymbol{x}_k$ for $\left(\boldsymbol{x}_0\right)_k$, the k-th column of $\boldsymbol{X}_0$. The function $g\left(\boldsymbol{w}\right)$ means $g\left(\boldsymbol{w};\boldsymbol{X}_0\right)$. We need the following lemmas to prove the Lipschitz results.

Lemma 9.5 *Suppose that $\varphi_1:U\to V$ is an L-Lipschitz map from a normed space U to a normed space V, and that $\varphi_2:V\to W$ is an L'-Lipschitz map from V to a normed space W. Then the composition $\varphi_2\circ\varphi_1:U\to W$ is LL'-Lipschitz.*

Lemma 9.6 *Fix any $\mathcal{D}\subseteq\mathbb{R}^{n-1}$. Let $g_1,g_2:\mathcal{D}\to\mathbb{R}$, and assume that g_1 is L_1-Lipschitz, and g_2 is L_2-Lipschitz, and that g_1 and g_2 are bounded over $\mathcal{D}$, i.e., $|g_1(\boldsymbol{x})|\leq M_1$ and $|g_2(\boldsymbol{x})|\leq M_2$ for all $x\in\mathcal{D}$ with some constants $M_1>0$ and $M_2>0$. Then the function $h(\boldsymbol{x})=g_1(\boldsymbol{x})g_2(\boldsymbol{x})$ is L-Lipschitz, with*

$$L = M_1L_2+M_2L_1.$$

Lemma 9.7 *For every $\boldsymbol{w},\boldsymbol{w}'\in\Gamma$, and every fixed $\boldsymbol{x}$, we have*

$$\begin{aligned}\left|\dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}\right)-\dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w}')^*\boldsymbol{x}\right)\right| &\leq \frac{2\sqrt{n}}{\mu}\left\|\boldsymbol{x}\right\|\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|,\\ \left|\ddot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}\right)-\ddot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w}')^*\boldsymbol{x}\right)\right| &\leq \frac{4\sqrt{n}}{\mu^2}\left\|\boldsymbol{x}\right\|\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|.\end{aligned}$$

Proof We have

$$\left|q_n\left(\boldsymbol{w}\right)-q_n\left(\boldsymbol{w}'\right)\right|=\left|\sqrt{1-\left\|\boldsymbol{w}\right\|^2}-\sqrt{1-\left\|\boldsymbol{w}'\right\|^2}\right|=\frac{\left\|\boldsymbol{w}+\boldsymbol{w}'\right\|\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|}{\sqrt{1-\left\|\boldsymbol{w}\right\|^2}+\sqrt{1-\left\|\boldsymbol{w}'\right\|^2}}$$

$$\leq \frac{\max\left(\left\|\boldsymbol{w}\right\|, \left\|\boldsymbol{w}'\right\|\right)}{\min\left(q_n\left(\boldsymbol{w}\right), q_n\left(\boldsymbol{w}'\right)\right)} \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|.$$

Hence it holds that

$$\left\|\boldsymbol{q}\left(\boldsymbol{w}\right) - \boldsymbol{q}\left(\boldsymbol{w}'\right)\right\|^2 = \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|^2 + \left|q_n\left(\boldsymbol{w}\right) - q_n\left(\boldsymbol{w}'\right)\right|^2 \leq \left(1 + \frac{\max\left(\left\|\boldsymbol{w}\right\|^2, \left\|\boldsymbol{w}'\right\|^2\right)}{\min\left(q_n^2\left(\boldsymbol{w}\right), q_n^2\left(\boldsymbol{w}'\right)\right)}\right) \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|^2$$

$$= \frac{1}{\min\left(q_n^2\left(\boldsymbol{w}\right), q_n^2\left(\boldsymbol{w}'\right)\right)} \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|^2 \leq 4n \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|^2,$$

where we have used the fact $q_n\left(\boldsymbol{w}\right) \geq \frac{1}{2\sqrt{n}}$ to get the final result. Hence the mapping $\boldsymbol{w} \mapsto \boldsymbol{q}(\boldsymbol{w})$ is $2\sqrt{n}$-Lipschitz over Γ. Moreover it is easy to see $\boldsymbol{q} \mapsto \boldsymbol{q}^* \boldsymbol{x}$ is $\left\|\boldsymbol{x}\right\|_2$-Lipschitz. By Lemma B.1 and the composition rule in Lemma 9.5, we obtain the desired claims. ■

> **Lemma 9.8** *For any fixed $\boldsymbol{x}$, consider the function*
>
> $$t_{\boldsymbol{x}}(\boldsymbol{w}) \doteq \frac{\boldsymbol{w}^* \overline{\boldsymbol{x}}}{\left\|\boldsymbol{w}\right\|} - \frac{x_n}{q_n(\boldsymbol{w})} \left\|\boldsymbol{w}\right\|$$
>
> *defined over $\boldsymbol{w} \in \Gamma$. Then, for all $\boldsymbol{w}, \boldsymbol{w}'$ in Γ such that $\left\|\boldsymbol{w}\right\| \geq r$ and $\left\|\boldsymbol{w}'\right\| \geq r$ for some constant $r \in (0,1)$, it holds that*
>
> $$\begin{aligned} \left|t_{\boldsymbol{x}}(\boldsymbol{w}) - t_{\boldsymbol{x}}(\boldsymbol{w}')\right| &\leq 2\left(\frac{\left\|\boldsymbol{x}\right\|}{r} + 4n^{3/2}\left\|\boldsymbol{x}\right\|_\infty\right)\left\|\boldsymbol{w} - \boldsymbol{w}'\right\|, \\ \left|t_{\boldsymbol{x}}(\boldsymbol{w})\right| &\leq 2\sqrt{n}\left\|\boldsymbol{x}\right\|, \\ \left|t_{\boldsymbol{x}}^2(\boldsymbol{w}) - t_{\boldsymbol{x}}^2(\boldsymbol{w}')\right| &\leq 8\sqrt{n}\left\|\boldsymbol{x}\right\|\left(\frac{\left\|\boldsymbol{x}\right\|}{r} + 4n^{3/2}\left\|\boldsymbol{x}\right\|_\infty\right)\left\|\boldsymbol{w} - \boldsymbol{w}'\right\|, \\ \left|t_{\boldsymbol{x}}^2(\boldsymbol{w})\right| &\leq 4n\left\|\boldsymbol{x}\right\|^2. \end{aligned}$$

Proof First of all, we have

$$\left|t_{\boldsymbol{x}}(\boldsymbol{w})\right| = \left[\frac{\boldsymbol{w}^*}{\left\|\boldsymbol{w}\right\|}, -\frac{\left\|\boldsymbol{w}\right\|}{q_n(\boldsymbol{w})}\right]\boldsymbol{x} \leq \left\|\boldsymbol{x}\right\|\left(1 + \frac{\left\|\boldsymbol{w}\right\|^2}{q_n^2(\boldsymbol{w})}\right)^{1/2} = \frac{\left\|\boldsymbol{x}\right\|}{\left|q_n(\boldsymbol{w})\right|} \leq 2\sqrt{n}\left\|\boldsymbol{x}\right\|,$$

where we have used the assumption that $q_n\left(\boldsymbol{w}\right) \geq \frac{1}{2\sqrt{n}}$ to simplify the final result. The claim about $\left|t_{\boldsymbol{x}}^2\left(\boldsymbol{w}\right)\right|$ follows immediately. Now

$$\left|t_{\boldsymbol{x}}(\boldsymbol{w}) - t_{\boldsymbol{x}}(\boldsymbol{w}')\right| \leq \left|\left(\frac{\boldsymbol{w}}{\left\|\boldsymbol{w}\right\|} - \frac{\boldsymbol{w}'}{\left\|\boldsymbol{w}'\right\|}\right)^* \overline{\boldsymbol{x}}\right| + \left|x_n\right|\left|\frac{\left\|\boldsymbol{w}\right\|}{q_n(\boldsymbol{w})} - \frac{\left\|\boldsymbol{w}'\right\|}{q_n(\boldsymbol{w}')}\right|.$$

Moreover we have

$$\left|\left(\frac{\boldsymbol{w}}{\left\|\boldsymbol{w}\right\|} - \frac{\boldsymbol{w}'}{\left\|\boldsymbol{w}'\right\|}\right)^* \overline{\boldsymbol{x}}\right| \leq \left\|\overline{\boldsymbol{x}}\right\|\left\|\frac{\boldsymbol{w}}{\left\|\boldsymbol{w}\right\|} - \frac{\boldsymbol{w}'}{\left\|\boldsymbol{w}'\right\|}\right\| \leq \left\|\boldsymbol{x}\right\|\frac{\left\|\boldsymbol{w} - \boldsymbol{w}'\right\|\left\|\boldsymbol{w}'\right\| + \left\|\boldsymbol{w}'\right\|\left|\left\|\boldsymbol{w}\right\| - \left\|\boldsymbol{w}'\right\|\right|}{\left\|\boldsymbol{w}\right\|\left\|\boldsymbol{w}'\right\|}$$

$$\leq \frac{2\left\|\boldsymbol{x}\right\|}{r}\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|,$$

where we have used the assumption that $\|\boldsymbol{w}\| \geq r$ to simplify the result. Noticing that $t \mapsto t/\sqrt{1-t^2}$ is continuous over $[a,b]$ and differentiable over (a,b) for any $0 < a < b < 1$, by mean value theorem,

$$\left|\frac{\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})} - \frac{\|\boldsymbol{w}'\|}{q_n(\boldsymbol{w}')}\right| \leq \sup_{\boldsymbol{w}\,\in\,\Gamma} \frac{1}{\left(1-\|\boldsymbol{w}\|^2\right)^{3/2}}\left\|\boldsymbol{w}-\boldsymbol{w}'\right\| \leq 8n^{3/2}\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|,$$

where we have again used the assumption that $q_n\left(\boldsymbol{w}\right) \geq \frac{1}{2\sqrt{n}}$ to simplify the last result. Collecting the above estimates, we obtain

$$\left|t_{\boldsymbol{x}}(\boldsymbol{w}) - t_{\boldsymbol{x}}(\boldsymbol{w}')\right| \leq \left(2\frac{\|\boldsymbol{x}\|}{r} + 8n^{3/2}\left\|\boldsymbol{x}\right\|_\infty\right)\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|,$$

as desired. For the last one, we have

$$\begin{aligned}\left|t_{\boldsymbol{x}}^2(\boldsymbol{w}) - t_{\boldsymbol{x}}^2(\boldsymbol{w}')\right| &= \left|t_{\boldsymbol{x}}(\boldsymbol{w}) - t_{\boldsymbol{x}}(\boldsymbol{w}')\right|\left|t_{\boldsymbol{x}}(\boldsymbol{w}) + t_{\boldsymbol{x}}(\boldsymbol{w}')\right| \\ &\leq 2\sup_{\boldsymbol{s}\,\in\,\Gamma}\left|t_{\boldsymbol{x}}(\boldsymbol{s})\right|\left|t_{\boldsymbol{x}}(\boldsymbol{w}) - t_{\boldsymbol{x}}(\boldsymbol{w}')\right|,\end{aligned}$$

leading to the claimed result once we substitute estimates of the involved quantities. ■

Lemma 9.9 *For any fixed $\boldsymbol{x}$, consider the function*

$$\boldsymbol{\Phi}_{\boldsymbol{x}}(\boldsymbol{w}) = \frac{x_n}{q_n(\boldsymbol{w})}\boldsymbol{I} + \frac{x_n}{q_n^3(\boldsymbol{w})}\boldsymbol{w}\boldsymbol{w}^*$$

defined over $\boldsymbol{w} \in \Gamma$. Then, for all $\boldsymbol{w}, \boldsymbol{w}' \in \Gamma$ such that $\|\boldsymbol{w}\| < r$ and $\|\boldsymbol{w}'\| < r$ with some constant $r \in \left(0, \frac{1}{2}\right)$, it holds that

$$\begin{aligned}\left\|\boldsymbol{\Phi}_{\boldsymbol{x}}(\boldsymbol{w})\right\| &\leq 2\left\|\boldsymbol{x}\right\|_\infty, \\ \left\|\boldsymbol{\Phi}_{\boldsymbol{x}}(\boldsymbol{w}) - \boldsymbol{\Phi}_{\boldsymbol{x}}(\boldsymbol{w}')\right\| &\leq 4\left\|\boldsymbol{x}\right\|_\infty\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|.\end{aligned}$$

Proof Simple calculation shows

$$\left\|\boldsymbol{\Phi}_{\boldsymbol{x}}(\boldsymbol{w})\right\| \leq \left\|\boldsymbol{x}\right\|_\infty\left(\frac{1}{q_n(\boldsymbol{w})} + \frac{\|\boldsymbol{w}\|^2}{q_n^3(\boldsymbol{w})}\right) = \frac{\|\boldsymbol{x}\|_\infty}{q_n^3(\boldsymbol{w})} \leq \frac{\|\boldsymbol{x}\|_\infty}{(1-r^2)^{3/2}} \leq 2\left\|\boldsymbol{x}\right\|_\infty.$$

For the second one, we have

$$\left\|\boldsymbol{\Phi}_{\boldsymbol{x}}(\boldsymbol{w}) - \boldsymbol{\Phi}_{\boldsymbol{x}}(\boldsymbol{w}')\right\| \leq \left\|\boldsymbol{x}\right\|_\infty\left\|\frac{1}{q_n(\boldsymbol{w})}\boldsymbol{I} + \frac{1}{q_n^3(\boldsymbol{w})}\boldsymbol{w}\boldsymbol{w}^* - \frac{1}{q_n(\boldsymbol{w}')}\boldsymbol{I} - \frac{1}{q_n^3(\boldsymbol{w}')}\boldsymbol{w}'(\boldsymbol{w}')^*\right\|$$

$$\leq \|\boldsymbol{x}\|_{\infty} \left(\left| \frac{1}{q_n(\boldsymbol{w})} - \frac{1}{q_n(\boldsymbol{w}')} \right| + \left| \frac{\|\boldsymbol{w}\|^2}{q_n^3(\boldsymbol{w})} - \frac{\|\boldsymbol{w}'\|^2}{q_n^3(\boldsymbol{w}')} \right| \right).$$

Now

$$\left| \frac{1}{q_n(\boldsymbol{w})} - \frac{1}{q_n(\boldsymbol{w}')} \right| = \frac{|q_n(\boldsymbol{w}) - q_n(\boldsymbol{w}')|}{q_n(\boldsymbol{w}) q_n(\boldsymbol{w}')} \leq \frac{\max(\|\boldsymbol{w}\|, \|\boldsymbol{w}'\|)}{\min(q_n^3(\boldsymbol{w}), q_n^3(\boldsymbol{w}'))} \|\boldsymbol{w} - \boldsymbol{w}'\| \leq \frac{4}{3\sqrt{3}} \|\boldsymbol{w} - \boldsymbol{w}'\|,$$

where we have applied the estimate for $|q_n(\boldsymbol{w}) - q_n(\boldsymbol{w}')|$ as established in Lemma 9.7 and also used $\|\boldsymbol{w}\| \leq 1/2$ and $\|\boldsymbol{w}'\| \leq 1/2$ to simplify the above result. Further noticing $t \mapsto t^2/\left(1-t^2\right)^{3/2}$ is differentiable over $t \in (0,1)$, we apply the mean value theorem and obtain

$$\left| \frac{\|\boldsymbol{w}\|^2}{q_n^3(\boldsymbol{w})} - \frac{\|\boldsymbol{w}'\|^2}{q_n^3(\boldsymbol{w}')} \right| \leq \sup_{\boldsymbol{s} \in \Gamma, \|\boldsymbol{s}\| \leq r < \frac{1}{2}} \frac{\|\boldsymbol{s}\|^3 + 2\|\boldsymbol{s}\|}{\left(1 - \|\boldsymbol{s}\|^2\right)^{5/2}} \|\boldsymbol{w} - \boldsymbol{w}'\| \leq \frac{4}{\sqrt{3}} \|\boldsymbol{w} - \boldsymbol{w}'\|.$$

Combining the above estimates gives the claimed result. ■

Lemma 9.10 *For any fixed $\boldsymbol{x}$, consider the function*

$$\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w}) = \overline{\boldsymbol{x}} - \frac{x_n}{q_n(\boldsymbol{w})} \boldsymbol{w}$$

defined over $\boldsymbol{w} \in \Gamma$. Then, for all $\boldsymbol{w}, \boldsymbol{w}' \in \Gamma$ such that $\|\boldsymbol{w}\| \leq r$ and $\|\boldsymbol{w}'\| \leq r$ for some constant $r \in \left(0, \frac{1}{2}\right)$, it holds that

$$\begin{aligned} \|\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})^*\| &\leq 2n\|\boldsymbol{x}\|_{\infty}^2, \\ \|\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})^* - \boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w}')\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w}')^*\| &\leq 8\sqrt{2}\sqrt{n}\|\boldsymbol{x}\|_{\infty}^2 \|\boldsymbol{w} - \boldsymbol{w}'\|. \end{aligned}$$

Proof We have $\|\boldsymbol{w}\|^2 / q_n^2(\boldsymbol{w}) \leq 1/3$ when $\|\boldsymbol{w}\| \leq r < 1/2$, hence it holds that

$$\|\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})^*\| \leq \|\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})\|^2 \leq 2\|\overline{\boldsymbol{x}}\|^2 + 2x_n^2 \frac{\|\boldsymbol{w}\|}{q_n^2(\boldsymbol{w})} \leq 2n\|\boldsymbol{x}\|_{\infty}^2.$$

For the second, we first estimate

$$\begin{aligned} \|\boldsymbol{\zeta}(\boldsymbol{w}) - \boldsymbol{\zeta}(\boldsymbol{w}')\| = \left\| x_n \left(\frac{\boldsymbol{w}}{q_n(\boldsymbol{w})} - \frac{\boldsymbol{w}'}{q_n(\boldsymbol{w}')} \right) \right\| &\leq \|\boldsymbol{x}\|_{\infty} \left\| \frac{\boldsymbol{w}}{q_n(\boldsymbol{w})} - \frac{\boldsymbol{w}'}{q_n(\boldsymbol{w}')} \right\| \\ &\leq \|\boldsymbol{x}\|_{\infty} \left(\frac{1}{q_n(\boldsymbol{w})} \|\boldsymbol{w} - \boldsymbol{w}'\| + \|\boldsymbol{w}'\| \left| \frac{1}{q_n(\boldsymbol{w})} - \frac{1}{q_n(\boldsymbol{w}')} \right| \right) \\ &\leq \|\boldsymbol{x}\|_{\infty} \left(\frac{1}{q_n(\boldsymbol{w})} + \frac{\|\boldsymbol{w}'\|}{\min\{q_n^3(\boldsymbol{w}), q_n^3(\boldsymbol{w}')\}} \right) \|\boldsymbol{w} - \boldsymbol{w}'\| \\ &\leq \|\boldsymbol{x}\|_{\infty} \left(\frac{2}{\sqrt{3}} + \frac{4}{3\sqrt{3}} \right) \|\boldsymbol{w} - \boldsymbol{w}'\| \leq 4\|\boldsymbol{x}\|_{\infty} \|\boldsymbol{w} - \boldsymbol{w}'\|. \end{aligned}$$

Thus, we have

$$\begin{aligned}\left\|\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})^* - \boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w}')\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w}')^*\right\| &\leq \left\|\boldsymbol{\zeta}(\boldsymbol{w})\right\|\left\|\boldsymbol{\zeta}(\boldsymbol{w}) - \zeta(\boldsymbol{w}')\right\| + \left\|\boldsymbol{\zeta}(\boldsymbol{w}) - \zeta(\boldsymbol{w}')\right\|\left\|\boldsymbol{\zeta}(\boldsymbol{w}')\right\| \\ &\leq 8\sqrt{2}\sqrt{n}\left\|\boldsymbol{x}\right\|_\infty^2\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|,\end{aligned}$$

as desired. ■

Now, we are ready to prove all the Lipschitz propositions.

Proof [of Proposition 4.11] Let

$$F_k(\boldsymbol{w}) = \ddot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)t_{\boldsymbol{x}_k}^2(\boldsymbol{w}) + \dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)\frac{x_k(n)}{q_n^3(\boldsymbol{w})}.$$

Then, $\frac{1}{\|\boldsymbol{w}\|^2}\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w} = \frac{1}{p}\sum_{k=1}^p F_k(\boldsymbol{w})$. Noticing that $\ddot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)$ is bounded by $1/\mu$ and $\dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)$ is bounded by 1, both in magnitude. Applying Lemma 9.6, Lemma 9.7 and Lemma 9.8, we can see $F_k(\boldsymbol{w})$ is $L_\frown^k$-Lipschitz with

$$\begin{aligned}L_\frown^k &= 4n\left\|\boldsymbol{x}_k\right\|^2\frac{4\sqrt{n}}{\mu^2}\left\|\boldsymbol{x}_k\right\| + \frac{1}{\mu}8\sqrt{n}\left\|\boldsymbol{x}_k\right\|\left(\frac{\left\|\boldsymbol{x}_k\right\|}{r_\frown} + 4n^{3/2}\left\|\boldsymbol{x}_k\right\|_\infty\right) \\ &\quad + (2\sqrt{n})^3\left\|\boldsymbol{x}_k\right\|_\infty\frac{2\sqrt{n}}{\mu}\left\|\boldsymbol{x}_k\right\| + \sup_{r_\frown<a<\sqrt{\frac{2n-1}{2n}}}\frac{3}{(1-a^2)^{5/2}}\left\|\boldsymbol{x}_k\right\|_\infty \\ &= \frac{16n^{3/2}}{\mu^2}\left\|\boldsymbol{x}_k\right\|^3 + \frac{8\sqrt{n}}{\mu r_\frown}\left\|\boldsymbol{x}_k\right\|^2 + \frac{48n^2}{\mu}\left\|\boldsymbol{x}_k\right\|\left\|\boldsymbol{x}_k\right\|_\infty + 96n^{5/2}\left\|\boldsymbol{x}_k\right\|_\infty.\end{aligned}$$

Thus, $\frac{1}{\|\boldsymbol{w}\|_2}\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w}$ is $L_\frown$-Lipschitz with

$$L_\frown \leq \frac{1}{p}\sum_{k=1}^p L_\frown^k \leq \frac{16n^3}{\mu^2}\left\|\boldsymbol{X}\right\|_\infty^3 + \frac{8n^{3/2}}{\mu r_\frown}\left\|\boldsymbol{X}\right\|_\infty^2 + \frac{48n^{5/2}}{\mu}\left\|\boldsymbol{X}\right\|_\infty^2 + 96n^{5/2}\left\|\boldsymbol{X}\right\|_\infty,$$

as desired. ■

Proof [of Proposition 4.12] We have

$$\left\|\frac{\boldsymbol{w}^*}{\|\boldsymbol{w}\|}\nabla g(\boldsymbol{w}) - \frac{\boldsymbol{w}'^*}{\|\boldsymbol{w}'\|}\nabla g(\boldsymbol{w}')\right\| \leq \frac{1}{p}\sum_{k=1}^p\left\|\dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)t_{\boldsymbol{x}_k}(\boldsymbol{w}) - \dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w}')^*\boldsymbol{x}_k\right)t_{\boldsymbol{x}_k}(\boldsymbol{w}')\right\|$$

where $\dot{h}_\mu(t) = \tanh(t/\mu)$ is bounded by one in magnitude, and $t_{\boldsymbol{x}_k}(\boldsymbol{w})$ and $t_{\boldsymbol{x}_k'}(\boldsymbol{w})$ is defined as in Lemma 9.8. By Lemma 9.6, Lemma 9.7 and Lemma 9.8, we know that $\dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)t_{\boldsymbol{x}_k}(\boldsymbol{w})$ is L_k-Lipschitz with constant

$$L_k = \frac{2\left\|\boldsymbol{x}_k\right\|}{r_g} + 8n^{3/2}\left\|\boldsymbol{x}_k\right\|_\infty + \frac{4n}{\mu}\left\|\boldsymbol{x}_k\right\|^2.$$

Therefore, we have

$$\begin{aligned}\left\| \frac{\boldsymbol{w}^*}{\|\boldsymbol{w}\|}\nabla g(\boldsymbol{w}) - \frac{\boldsymbol{w}^*}{\|\boldsymbol{w}\|}\nabla g(\boldsymbol{w}')\right\| &\le \frac{1}{p}\sum_{k=1}^{p}\left(\frac{2\left\|\boldsymbol{x}_k\right\|}{r_g} + 8n^{3/2}\left\|\boldsymbol{x}_k\right\|_\infty + \frac{4n}{\mu}\left\|\boldsymbol{x}_k\right\|^2\right)\left\|\boldsymbol{w}-\boldsymbol{w}'\right\| \\ &\le \left(\frac{2\sqrt{n}}{r_g}\left\|\boldsymbol{X}\right\|_\infty + 8n^{3/2}\left\|\boldsymbol{X}\right\|_\infty + \frac{4n^2}{\mu}\left\|\boldsymbol{X}\right\|_\infty^2\right)\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|,\end{aligned}$$

as desired. ■

Proof [of Proposition 4.13] Let

$$\boldsymbol{F}_k(\boldsymbol{w}) = \ddot{h}_\mu(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k)\boldsymbol{\zeta}_k(\boldsymbol{w})\boldsymbol{\zeta}_k(\boldsymbol{w})^* - \dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)\boldsymbol{\Phi}_k(\boldsymbol{w})$$

with $\boldsymbol{\zeta}_k(\boldsymbol{w}) = \overline{\boldsymbol{x}}_k - \frac{x_k(n)}{q_n(\boldsymbol{w})}\boldsymbol{w}$ and $\boldsymbol{\Phi}_k(\boldsymbol{w}) = \frac{x_k(n)}{q_n(\boldsymbol{w})}\boldsymbol{I} + \frac{x_{n,k}}{q_n(\boldsymbol{w})}\boldsymbol{w}\boldsymbol{w}^*$. Then, $\nabla^2 g(\boldsymbol{w}) = \frac{1}{p}\sum_{k=1}^{p}\boldsymbol{F}_k(\boldsymbol{w})$. Using Lemma 9.6, Lemma 9.7, Lemma 9.9 and Lemma 9.10, and the facts that $\ddot{h}_\mu(t)$ is bounded by $1/\mu$ and that $\ddot{h}_\mu(t)$ is bounded by 1 in magnitude, we can see $\boldsymbol{F}_k(\boldsymbol{w})$ is $L_{\smile}^{k}$-Lipschitz continuous with

$$\begin{aligned} L_{\smile}^{k} &= \frac{1}{\mu}\times 8\sqrt{2}\sqrt{n}\left\|\boldsymbol{x}_k\right\|_\infty^2 + \frac{2\sqrt{n}}{\mu^2}\left\|\boldsymbol{x}_k\right\|\times 2n\left\|\boldsymbol{x}_k\right\|_\infty^2 + 4\left\|\boldsymbol{x}_k\right\|_\infty + \frac{2\sqrt{n}}{\mu}\left\|\boldsymbol{x}_k\right\|\times 2\left\|\boldsymbol{x}_k\right\|_\infty \\ &\le \frac{4n^{3/2}}{\mu^2}\left\|\boldsymbol{x}_k\right\|\left\|\boldsymbol{x}_k\right\|_\infty^2 + \frac{4\sqrt{n}}{\mu}\left\|\boldsymbol{x}_k\right\|\left\|\boldsymbol{x}_k\right\|_\infty + \frac{8\sqrt{2}\sqrt{n}}{\mu}\left\|\boldsymbol{x}_k\right\|_\infty^2 + 4\left\|\boldsymbol{x}_k\right\|_\infty.\end{aligned}$$

Thus, we have

$$L_{\smile} \le \frac{1}{p}\sum_{k=1}^{p}L_{\smile}^{k} \le \frac{4n^2}{\mu^2}\left\|\boldsymbol{X}\right\|_\infty^3 + \frac{4n}{\mu}\left\|\boldsymbol{X}\right\|_\infty^2 + \frac{8\sqrt{2}\sqrt{n}}{\mu}\left\|\boldsymbol{X}\right\|_\infty^2 + 8\left\|\boldsymbol{X}\right\|_\infty,$$

as desired. ■

9.2 Proofs of Theorem 4.1

To avoid clutter of notations, in this subsection we write $\boldsymbol{X}$ to mean $\boldsymbol{X}_0$; similarly $\boldsymbol{x}_k$ for $(\boldsymbol{x}_0)_k$, the k-th column of $\boldsymbol{X}_0$. The function $g(\boldsymbol{w})$ means $g(\boldsymbol{w};\boldsymbol{X}_0)$. Before proving Theorem 4.1, we record one useful lemma.

> **Lemma 9.11** *For any $\theta \in (0,1)$, consider the random matrix $\boldsymbol{X} \in \mathbb{R}^{n_1\times n_2}$ with $\boldsymbol{X} \sim_{i.i.d.} \mathrm{BG}(\theta)$. Define the event $\mathcal{E}_\infty \doteq \left\{1 \le \left\|\boldsymbol{X}\right\|_\infty \le 4\sqrt{\log(np)}\right\}$. It holds that*
>
> $$\mathbb{P}\left[\mathcal{E}_\infty^c\right] \le \theta\,(np)^{-7} + \exp\left(-0.3\theta np\right).$$

Proof See Section B.2.4 on Page 229. ■

For convenience, we define three regions covering the whole range of $\boldsymbol{w}$:

$$R_1 \doteq \left\{ \boldsymbol{w} : \|\boldsymbol{w}\| \le \frac{\mu}{4\sqrt{2}} \right\}, \qquad R_2 \doteq \left\{ \boldsymbol{w} : \frac{\mu}{4\sqrt{2}} \le \|\boldsymbol{w}\| \le \frac{1}{20\sqrt{5}} \right\},$$
$$R_3 \doteq \left\{ \boldsymbol{w} : \frac{1}{20\sqrt{5}} \le \|\boldsymbol{w}\| \le \sqrt{\frac{4n-1}{4n}} \right\}.$$

Proof [of Theorem 4.1] We will first use covering argument and continuity to show that the random quantities of interest in R_1, R_2, and R_3 concentrate uniformly around their expectations. Then, we will show that the only local minimizer is next to $\mathbf{0}$ and is contained in R_1.

Strong convexity in region R_1. Proposition 4.7 shows that for any $\boldsymbol{w} \in R_1$, $\mathbb{E}\left[\nabla^2 g(\boldsymbol{w})\right] \succeq \frac{c_1\theta}{\mu}\boldsymbol{I}$. For any $\varepsilon \in (0, \mu/\left(4\sqrt{2}\right))$, R_1 has an ε-net N_1 of size at most $(3\mu/\left(4\sqrt{2}\varepsilon\right))^n$. On $\mathcal{E}_\infty$, $\nabla^2 g$ is

$$L_1 \doteq \frac{C_2 n^2}{\mu^2}\log^{3/2}(np)$$

Lipschitz by Proposition 4.13. Set $\varepsilon = \frac{c_1\theta}{3\mu L_1}$, so

$$\#N_1 \le \exp\left(2n\log\left(\frac{C_3 n\log(np)}{\theta}\right)\right).$$

Let $\mathcal{E}_1$ denote the event

$$\mathcal{E}_1 = \left\{ \max_{\boldsymbol{w}\in N_1} \left\|\nabla^2 g(\boldsymbol{w}) - \mathbb{E}\left[\nabla^2 g(\boldsymbol{w})\right]\right\| \le \frac{c_1\theta}{3\mu} \right\}.$$

On $\mathcal{E}_1 \cap \mathcal{E}_\infty$,

$$\sup_{\|\boldsymbol{w}\|\le\mu/\left(4\sqrt{2}\right)} \left\|\nabla^2 g(\boldsymbol{w}) - \mathbb{E}\left[\nabla^2 g(\boldsymbol{w})\right]\right\| \ \le\ \frac{2c_1\theta}{3\mu},$$

and so on $\mathcal{E}_1 \cap \mathcal{E}_\infty$, (4.1.2) holds for any constant $c_\star \le c_1/3$. Setting $t = c_1\theta/3\mu$ in Proposition 4.10, we obtain that for any fixed $\boldsymbol{w}$,

$$\mathbb{P}\left[\left\|\nabla^2 g(\boldsymbol{w}) - \mathbb{E}\left[\nabla^2 g(\boldsymbol{w})\right]\right\| \ge \frac{c_1\theta}{3\mu}\right] \le 4n\exp\left(-\frac{c_4 p\theta^2}{n^2}\right).$$

Taking a union bound, we obtain that

$$\mathbb{P}\left[\mathcal{E}_1^c\right] \ \le\ 4n\exp\left(-\frac{c_4 p\theta^2}{n^2} + C_5 n\log(n) + C_5 n\log\log(p)\right).$$

Large gradient in region R_2. Similarly, for the gradient quantity, for $\boldsymbol{w} \in R_2$, Proposition 4.6 shows that

$$\mathbb{E}\left[\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|}\right] \ \ge\ c_6\theta.$$

Moreover, on $\mathcal{E}_\infty$, $\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|}$ is

$$L_2 \doteq \frac{C_7 n^2}{\mu} \log(np)$$

Lipschitz by Proposition 4.12. For any $\varepsilon < \frac{1}{20\sqrt{5}}$, the set R_2 has an ε-net N_2 of size at most $\left(\frac{3}{20\varepsilon\sqrt{5}}\right)^n$. Set $\varepsilon = \frac{c_6\theta}{3L_2}$, so

$$\# N_2 \;\le\; \exp\left(n \log\left(\frac{C_8 n^2 \log(np)}{\theta\mu} \right) \right).$$

Let $\mathcal{E}_2$ denote the event

$$\mathcal{E}_2 = \left\{ \max_{\boldsymbol{w} \in N_2} \left| \frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} \right] \right| \;\le\; \frac{c_6\theta}{3} \right\}.$$

On $\mathcal{E}_2 \cap \mathcal{E}_\infty$,

$$\sup_{\boldsymbol{w} \in R_2} \left| \frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} \right] \right| \;\le\; \frac{2c_6\theta}{3}, \tag{9.2.1}$$

and so on $\mathcal{E}_2 \cap \mathcal{E}_\infty$, (4.1.3) holds for any constant $c_\star \le c_6/3$. Setting $t = c_6\theta/3$ in Proposition 4.9, we obtain that for any fixed $\boldsymbol{w} \in R_2$,

$$\mathbb{P}\left[\left| \frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} \right] \right| \right] \;\le\; 2\exp\left(-\frac{c_9 p \theta^2}{n} \right),$$

and so

$$\mathbb{P}\left[\mathcal{E}_2^c\right] \;\le\; 2\exp\left(-\frac{c_9 p\theta^2}{n} + n\log\left(\frac{C_8 n^2 \log(np)}{\theta\mu} \right) \right). \tag{9.2.2}$$

Existence of negative curvature direction in R_3. Finally, for any $\boldsymbol{w} \in R_3$, Proposition 4.5 shows that

$$\mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} \right] \;\le\; -c_9\theta.$$

On $\mathcal{E}_\infty$, $\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2}$ is

$$L_3 = \frac{C_{10} n^3}{\mu^2} \log^{3/2}(np)$$

Lipschitz by Proposition 4.11. As above, for any $\varepsilon \le \sqrt{\frac{4n-1}{4n}}$, R_3 has an ε-net N_3 of size at most $(3/\varepsilon)^n$. Set $\varepsilon = c_9\theta/3L_3$. Then

$$\# N_3 \;\le\; \exp\left(n\log\left(\frac{C_{11} n^3 \log^{3/2}(np)}{\theta\mu^2} \right) \right).$$

Let $\mathcal{E}_3$ denote the event

$$\mathcal{E}_3 = \left\{ \max_{\boldsymbol{w} \in N_3} \left| \frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} \right] \right| \le \frac{c_9\theta}{3} \right\}$$

On $\mathcal{E}_3 \cap \mathcal{E}_\infty$,

$$\sup_{\boldsymbol{w}\in R_3} \left| \frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} \right] \right| \;\le\; \frac{2c_9\theta}{3},$$

and (4.1.4) holds with any constant $c_\star < c_9/3$. Setting $t = c_9\theta/3$ in Proposition 4.8 and taking a union bound, we obtain

$$\mathbb{P}\left[\mathcal{E}_3^c\right] \;\le\; 4\exp\left(-\frac{c_{12} p \mu^2 \theta^2}{n^2} + n \log\left(\frac{C_{11} n^3 \log^{3/2}(np)}{\theta \mu^2} \right) \right).$$

The unique local minimizer located near 0. Let $\mathcal{E}_g$ be the event that the bounds (4.1.2)-(4.1.4) hold. On $\mathcal{E}_g$, the function g is $\frac{c_\star\theta}{\mu}$-strongly convex over $R_1 = \left\{ \boldsymbol{w} \mid \|\boldsymbol{w}\| \le \mu/\left(4\sqrt{2}\right) \right\}$. This implies that f has at most one local minimum on R_1. It also implies that for any $\boldsymbol{w} \in R_1$,

$$g(\boldsymbol{w}) \ge g(\mathbf{0}) + \langle \nabla g(\mathbf{0}), \boldsymbol{w} \rangle + \frac{c\theta}{2\mu} \|\boldsymbol{w}\|^2 \ge g(\mathbf{0}) - \|\boldsymbol{w}\| \|\nabla g(\mathbf{0})\| + \frac{c_\star\theta}{2\mu} \|\boldsymbol{w}\|^2.$$

So, if $g(\boldsymbol{w}) \le g(\mathbf{0})$, we necessarily have

$$\|\boldsymbol{w}\| \;\le\; \frac{2\mu}{c_\star\theta} \|\nabla g(\mathbf{0})\|.$$

Suppose that

$$\|\nabla g(\mathbf{0})\| \le \frac{c_\star\theta}{32}. \tag{9.2.3}$$

Then $g(\boldsymbol{w}) \le g(\mathbf{0})$ implies that $\|\boldsymbol{w}\| \le \mu/16$. By Wierstrass's theorem, $g(\boldsymbol{w})$ has at least one minimizer $\boldsymbol{w}_\star$ over the compact set $S = \{\boldsymbol{w} \mid \|\boldsymbol{w}\| \le \mu/10\}$. By the above reasoning, $\|\boldsymbol{w}_\star\| \le \mu/16$, and hence $\boldsymbol{w}_\star$ does not lie on the boundary of S. This implies that $\boldsymbol{w}_\star$ is a local minimizer of g. Moreover, as above,

$$\|\boldsymbol{w}_\star\| \;\le\; \frac{2\mu}{c_\star\theta} \|\nabla g(\mathbf{0})\|.$$

We now use the vector Bernstein inequality to show that with our choice of p, (9.2.3) is satisfied with high probability. Notice that

$$\nabla g(\mathbf{0}) = \frac{1}{p} \sum_{i=1}^{p} \dot{h}_\mu(x_i(n)) \overline{\boldsymbol{x}}_i,$$

and $\dot{h}_\mu$ is bounded by one in magnitude, so for any integer $m \ge 2$,

$$\mathbb{E}\left[\left\| \dot{h}_\mu(x_i(n)) \overline{\boldsymbol{x}}_i \right\|^m \right] \le \mathbb{E}\left[\|\boldsymbol{x}_i\|^m \right] \le \mathbb{E}_{Z\sim\chi(n)}\left[Z^m \right] \le m! n^{m/2},$$

where we have applied the moment estimate for the $\chi(n)$ distribution shown in Lemma B.8. Applying the vector Bernstein inequality in Corollary A.3 with $R = \sqrt{n}$ and $\sigma^2 = 2n$, we obtain

$$\mathbb{P}\left[\left\|\nabla g(\mathbf{0})\right\| \;\geq\; t\right] \;\leq\; 2(n+1)\exp\left(-\frac{pt^2}{4n+2\sqrt{n}t}\right)$$

for all $t > 0$. Using this inequality, it is not difficult to show that there exist constants $C_{13}, C_{14} > 0$ such that when $p \geq C_{13} n \log n$, with probability at least $1 - 4np^{-10}$,

$$\left\|\nabla g(\mathbf{0})\right\| \;\leq\; C_3\sqrt{\frac{n\log p}{p}}. \tag{9.2.4}$$

When $\frac{p}{\log p} \geq \frac{C_{14} n}{\theta^2}$, for appropriately large C_{14}, (9.2.4) implies (9.2.3). Summing up failure probabilities completes the proof. ■

9.3 Proofs for Section 4.3 and Theorem 4.3

Proof [of Lemma 4.14] By the generative model,

$$\overline{\boldsymbol{Y}} = \left(\frac{1}{p\theta}\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y} = \left(\frac{1}{p\theta}\boldsymbol{A}_0\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0\boldsymbol{X}_0.$$

Since $\mathbb{E}\left[\boldsymbol{X}_0\boldsymbol{X}_0^*/\left(p\theta\right)\right] = \boldsymbol{I}$, we will compare $\left(\frac{1}{p\theta}\boldsymbol{A}_0\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0$ with $\left(\boldsymbol{A}_0\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0 = \boldsymbol{U}\boldsymbol{V}^*$. By Lemma B.11, we have

$$\begin{aligned}
&\left\|\left(\frac{1}{p\theta}\boldsymbol{A}_0\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0 - \left(\boldsymbol{A}_0\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0\right\| \\
\leq\;& \left\|\boldsymbol{A}_0\right\|\left\|\left(\frac{1}{p\theta}\boldsymbol{A}_0\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{A}_0^*\right)^{-1/2} - \left(\boldsymbol{A}_0\boldsymbol{A}_0^*\right)^{-1/2}\right\| \\
\leq\;& \left\|\boldsymbol{A}_0\right\|\frac{2\left\|\boldsymbol{A}_0\right\|^3}{\sigma_{\min}^4\left(\boldsymbol{A}_0\right)}\left\|\frac{1}{p\theta}\boldsymbol{X}_0\boldsymbol{X}_0^* - \boldsymbol{I}\right\| = 2\kappa^4\left(\boldsymbol{A}_0\right)\left\|\frac{1}{p\theta}\boldsymbol{X}_0\boldsymbol{X}_0^* - \boldsymbol{I}\right\|
\end{aligned}$$

provided

$$\left\|\boldsymbol{A}_0\right\|^2\left\|\frac{1}{p\theta}\boldsymbol{X}_0\boldsymbol{X}_0^* - \boldsymbol{I}\right\| \leq \frac{\sigma_{\min}^2\left(\boldsymbol{A}_0\right)}{2} \Longleftrightarrow \left\|\frac{1}{p\theta}\boldsymbol{X}_0\boldsymbol{X}_0^* - \boldsymbol{I}\right\| \leq \frac{1}{2\kappa^2\left(\boldsymbol{A}_0\right)}.$$

On the other hand, by Lemma B.12, when $p \geq C_1 n^2 \log n$ for some large constant C_1, $\left\|\frac{1}{p\theta}\boldsymbol{X}_0\boldsymbol{X}_0^* - \boldsymbol{I}\right\| \leq 10\sqrt{\frac{\theta n \log p}{p}}$ with probability at least $1 - p^{-8}$. Thus, when $p \geq C_2\kappa^4\left(\boldsymbol{A}_0\right)\theta n^2\log(n\theta\kappa\left(\boldsymbol{A}_0\right))$,

$$\left\|\left(\frac{1}{p\theta}\boldsymbol{A}_0\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0 - \left(\boldsymbol{A}_0\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0\right\| \leq 20\kappa^4\left(\boldsymbol{A}_0\right)\sqrt{\frac{\theta n\log p}{p}},$$

as desired. ■

Proof [of Lemma 4.15] To avoid clutter in notation, we write $\boldsymbol{X}$ to mean $\boldsymbol{X}_0$, and $\boldsymbol{x}_k$ to mean $(\boldsymbol{x}_0)_k$ in this proof. We also let $\widetilde{\boldsymbol{Y}} \doteq \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0$. Note the Jacobian matrix for the mapping $\boldsymbol{q}(\boldsymbol{w})$ is $\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w}) = \left[\boldsymbol{I}, -\boldsymbol{w}/\sqrt{1-\|\boldsymbol{w}\|^2}\right]$. Hence for any vector $\boldsymbol{z} \in \mathbb{R}^n$ and all $\boldsymbol{w} \in \Gamma$,

$$\|\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\boldsymbol{z}\| \le \sqrt{n-1}\,\|\boldsymbol{z}\|_\infty + \frac{\|\boldsymbol{w}\|}{\sqrt{1-\|\boldsymbol{w}\|^2}}\|\boldsymbol{z}\|_\infty \le 3\sqrt{n}\,\|\boldsymbol{z}\|_\infty .$$

Now we have

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{w}} g\left(\boldsymbol{w};\widetilde{\boldsymbol{Y}}\right) - \nabla_{\boldsymbol{w}} g(\boldsymbol{w};\boldsymbol{X})\right\| \\
={}& \left\|\frac{1}{p}\sum_{k=1}^{p}\dot{h}_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k + \boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{\Xi}}\boldsymbol{x}_k\right)\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\left(\boldsymbol{x}_k + \widetilde{\boldsymbol{\Xi}}\boldsymbol{x}_k\right) - \frac{1}{p}\sum_{k=1}^{p}\dot{h}_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k\right)\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\boldsymbol{x}_k\right\| \\
\le{}& \left\|\frac{1}{p}\sum_{k=1}^{p}\dot{h}_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k + \boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{\Xi}}\boldsymbol{x}_k\right)\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\left(\boldsymbol{x}_k + \widetilde{\boldsymbol{\Xi}}\boldsymbol{x}_k - \boldsymbol{x}_k\right)\right\| \\
&\quad + \left\|\frac{1}{p}\sum_{k=1}^{p}\left[\dot{h}_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k + \boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{\Xi}}\boldsymbol{x}_k\right) - \dot{h}_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k\right)\right]\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\boldsymbol{x}_k\right\| \\
\le{}& \left\|\widetilde{\boldsymbol{\Xi}}\right\|\left(\max_t \dot{h}_\mu(t)\,3n\,\|\boldsymbol{X}\|_\infty + L_{\dot{h}_\mu}3n\,\|\boldsymbol{X}\|_\infty^2\right),
\end{aligned}$$

where $L_{\dot{h}_\mu}$ denotes the Lipschitz constant for $\dot{h}_\mu(\cdot)$. Similarly, suppose $\left\|\widetilde{\boldsymbol{\Xi}}\right\| \le \frac{1}{2n}$, and also notice that

$$\left\|\frac{\boldsymbol{I}}{q_n(\boldsymbol{w})} + \frac{\boldsymbol{w}\boldsymbol{w}^*}{q_n^3(\boldsymbol{w})}\right\| \le \frac{1}{q_n(\boldsymbol{w})} + \frac{\|\boldsymbol{w}\|^2}{q_n^3(\boldsymbol{w})} = \frac{1}{q_n^3(\boldsymbol{w})} \le 2\sqrt{2}n^{3/2},$$

we obtain that

$$\begin{aligned}
&\left\|\nabla^2_{\boldsymbol{w}} g\left(\boldsymbol{w};\widetilde{\boldsymbol{Y}}\right) - \nabla^2_{\boldsymbol{w}} g(\boldsymbol{w};\boldsymbol{X})\right\| \\
\le{}& \left\|\frac{1}{p}\sum_{k=1}^{p}\left[\ddot{h}\left(\boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{y}}_k\right)\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\widetilde{\boldsymbol{y}}_k\widetilde{\boldsymbol{y}}_k^*\left(\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\right)^* - \ddot{h}\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k\right)\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\boldsymbol{x}_k\boldsymbol{x}_k^*\left(\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\right)^*\right]\right\| \\
&\quad + \left\|\frac{1}{p}\sum_{k=1}^{p}\left[\dot{h}\left(\boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{y}}_k\right)\left(\frac{\boldsymbol{I}}{q_n(\boldsymbol{w})} + \frac{\boldsymbol{w}\boldsymbol{w}^*}{q_n^3}\right)\widetilde{\boldsymbol{y}}_k(n) - \dot{h}\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k\right)\left(\frac{\boldsymbol{I}}{q_n(\boldsymbol{w})} + \frac{\boldsymbol{w}\boldsymbol{w}^*}{q_n^3}\right)\boldsymbol{x}_k(n)\right]\right\| \\
\le{}& \tfrac{45}{2}L_{\ddot{h}_\mu}n^{3/2}\|\boldsymbol{X}\|_\infty^3\left\|\widetilde{\boldsymbol{\Xi}}\right\| + \max_t \ddot{h}_\mu(t)\left(18n^{3/2}\|\boldsymbol{X}\|_\infty^2\left\|\widetilde{\boldsymbol{\Xi}}\right\| + 10n^2\|\boldsymbol{X}\|_\infty^2\left\|\widetilde{\boldsymbol{\Xi}}\right\|^2\right) \\
&\quad + 3\sqrt{2}L_{\dot{h}_\mu}n^2\left\|\widetilde{\boldsymbol{\Xi}}\right\|\|\boldsymbol{X}\|_\infty^2 + \max_t \dot{h}(t)\,2\sqrt{2}n^2\left\|\widetilde{\boldsymbol{\Xi}}\right\|\|\boldsymbol{X}\|_\infty,
\end{aligned}$$

where $L_{\ddot{h}_\mu}$ denotes the Lipschitz constant for $\ddot{h}_\mu\left(\cdot\right)$. Since

$$\max_t \dot{h}_\mu\left(t\right) \le 1, \quad \max_t \ddot{h}_\mu\left(t\right) \le \frac{1}{\mu}, \quad L_{h_\mu} \le 1, \quad L_{\dot{h}_\mu} \le \frac{1}{\mu}, \quad L_{\ddot{h}_\mu} \le \frac{2}{\mu^2},$$

and by Lemma 9.11, $\left\|\boldsymbol{X}\right\|_\infty \le 4\sqrt{\log\left(np\right)}$ with probability at least $1-\theta\left(np\right)^{-7}-\exp\left(-0.3\theta np\right)$, we obtain

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{w}} g\left(\boldsymbol{w}; \widetilde{\boldsymbol{Y}}\right) - \nabla_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\| &\le C_1 \frac{n}{\mu} \log\left(np\right) \left\|\widetilde{\boldsymbol{\Xi}}\right\|, \\
\left\|\nabla^2_{\boldsymbol{w}} g\left(\boldsymbol{w}; \widetilde{\boldsymbol{Y}}\right) - \nabla^2_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\| &\le C_2 \max\left\{\frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu}\right\} \log^{3/2}\left(np\right) \left\|\widetilde{\boldsymbol{\Xi}}\right\|
\end{aligned}$$

for constants $C_1, C_2 > 0$. ■

Proof [of Theorem 4.3] Assume the constant $c_\star$ as defined in Theorem 4.1. By Lemma 4.14, when

$$p \ge \frac{C_1}{c_\star^2 \theta} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8\left(\boldsymbol{A}_0\right) \log^4\left(\frac{\kappa\left(\boldsymbol{A}_0\right) n}{\mu\theta}\right),$$

the magnitude of the perturbation is bounded as

$$\left\|\widetilde{\boldsymbol{\Xi}}\right\| \le C_2 c_\star \theta \left(\max\left\{\frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu}\right\} \log^{3/2}\left(np\right)\right)^{-1},$$

where C_2 can be made arbitrarily small by making C_1 large. Combining this result with Lemma 4.15, we obtain that for all $\boldsymbol{w} \in \Gamma$,

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) - \nabla_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\| &\le \frac{c_\star\theta}{2} \\
\left\|\nabla^2_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) - \nabla^2_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\| &\le \frac{c_\star\theta}{2},
\end{aligned}$$

with probability at least $1-p^{-8}-\theta\left(np\right)^{-7}-\exp\left(-0.3\theta np\right)$. In view of (4.1.10) in Theorem 4.1, we have

$$\begin{aligned}
\frac{\boldsymbol{w}^* g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right)\boldsymbol{w}}{\left\|\boldsymbol{w}\right\|^2} &= \frac{\boldsymbol{w}^* g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)\boldsymbol{w}}{\left\|\boldsymbol{w}\right\|^2} + \frac{\boldsymbol{w}^* g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right)\boldsymbol{w}}{\left\|\boldsymbol{w}\right\|^2} - \frac{\boldsymbol{w}^* g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)\boldsymbol{w}}{\left\|\boldsymbol{w}\right\|^2} \\
&\le -c_\star\theta + \left\|\nabla^2_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) - \nabla^2_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\| \le -\frac{1}{2}c_\star\theta.
\end{aligned}$$

By similar arguments, we obtain (4.1.8) through (4.1.10) in Theorem 4.3.

To show the unique local minimizer over Γ is near $\mathbf{0}$, we note that (recall the last part of proof of Theorem 4.1 in Section 9.2) $g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right)$ being $\frac{c_\star\theta}{2\mu}$ strongly convex near $\mathbf{0}$ implies that

$$\left\|\boldsymbol{w}_\star\right\| \le \frac{4\mu}{c_\star\theta}\left\|\nabla g\left(\mathbf{0}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right)\right\|.$$

The above perturbation analysis implies there exists $C_3 > 0$ such that when

$$p \geq \frac{C_3}{c_\star^2 \theta} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8\left(\boldsymbol{A}_0\right) \log^4\left(\frac{\kappa\left(\boldsymbol{A}_0\right) n}{\mu\theta}\right),$$

it holds that

$$\left\|\nabla_{\boldsymbol{w}} g\left(\mathbf{0}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) - \nabla_{\boldsymbol{w}} g\left(\mathbf{0}; \boldsymbol{X}\right)\right\| \leq \frac{c_\star \theta}{400},$$

which in turn implies

$$\left\|\boldsymbol{w}_\star\right\| \leq \frac{4\mu}{c_\star \theta}\left\|\nabla g\left(\mathbf{0}; \boldsymbol{X}_0\right)\right\| + \frac{4\mu}{c_\star \theta}\frac{c_\star \theta}{400} \leq \frac{\mu}{8} + \frac{\mu}{100} < \frac{\mu}{7},$$

where we have recall the result that $\frac{2\mu}{c_\star \theta}\left\|\nabla g\left(\mathbf{0}; \boldsymbol{X}_0\right)\right\| \leq \mu/16$ from proof of Theorem 4.1. A simple union bound with careful bookkeeping gives the success probability. ■

Chapter 10

Proof of Convergence of the Trust-Region Algorithm

> I have had my results for a long time, but I do not yet know how to arrive at them.
>
> Karl Friedrich Gauss

10.1 Proof of Lemma 5.3

Proof Using the fact $\tanh(\cdot)$ and $1-\tanh^2(\cdot)$ are bounded by one in magnitude, by (5.3.1) and (5.3.2) we have

$$\begin{aligned}\left\|\nabla f\left(\boldsymbol{q}\right)\right\| &\le \frac{1}{p}\sum_{k=1}^{p}\left\|\boldsymbol{x}_k\right\| \le \sqrt{n}\left\|\boldsymbol{X}\right\|_{\infty}, \\ \left\|\nabla^2 f\left(\boldsymbol{q}\right)\right\| &\le \frac{1}{p}\sum_{k=1}^{p}\frac{1}{\mu}\left\|\boldsymbol{x}_k\right\|^2 \le \frac{n}{\mu}\left\|\boldsymbol{X}\right\|_{\infty}^2,\end{aligned}$$

for any $\boldsymbol{q}\in\mathbb{S}^{n-1}$. Moreover,

$$\begin{aligned}\sup_{\boldsymbol{q},\boldsymbol{q}'\in\mathbb{S}^{n-1},\boldsymbol{q}\neq\boldsymbol{q}'}\frac{\left\|\nabla f\left(\boldsymbol{q}\right)-\nabla f\left(\boldsymbol{q}'\right)\right\|}{\left\|\boldsymbol{q}-\boldsymbol{q}'\right\|} &\le \frac{1}{p}\sum_{k=1}^{p}\left\|\boldsymbol{x}_k\right\|\sup_{\boldsymbol{q},\boldsymbol{q}'\in\mathbb{S}^{n-1},\boldsymbol{q}\neq\boldsymbol{q}'}\frac{\left|\tanh\left(\frac{\boldsymbol{q}^*\boldsymbol{x}_k}{\mu}\right)-\tanh\left(\frac{\boldsymbol{q}'^*\boldsymbol{x}_k}{\mu}\right)\right|}{\left\|\boldsymbol{q}-\boldsymbol{q}'\right\|} \\ &\le \frac{1}{p}\sum_{k=1}^{p}\left\|\boldsymbol{x}_k\right\|\frac{\left\|\boldsymbol{x}_k\right\|}{\mu} \le \frac{n}{\mu}\left\|\boldsymbol{X}\right\|_{\infty}^2,\end{aligned}$$

where at the last line we have used the fact the mapping $\boldsymbol{q} \mapsto \boldsymbol{q}^* \boldsymbol{x}_k / \mu$ is $\|\boldsymbol{x}_k\| / \mu$ Lipschitz, and $\boldsymbol{x} \mapsto \tanh(x)$ is 1-Lipschitz, and the composition rule in Lemma 9.5. Similar argument yields the final bound. ■

10.2 Proof of Lemma 5.4

Proof Suppose we can establish

$$\left| f\left(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})\right) - \widehat{f}\left(\boldsymbol{q}, \boldsymbol{\delta}\right) \right| \;\le\; \frac{1}{6}\eta_f \left\|\boldsymbol{\delta}\right\|^3 .$$

Applying this twice we obtain

$$f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta}_\star)) \le \widehat{f}(\boldsymbol{q}, \boldsymbol{\delta}_\star) + \frac{1}{6}\eta_f \Delta^3 \le \widehat{f}(\boldsymbol{q}, \boldsymbol{\delta}) + \frac{1}{6}\eta_f \Delta^3 \le f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) + \frac{1}{3}\eta_f \Delta^3 \le f(\boldsymbol{q}) - s + \frac{1}{3}\eta_f \Delta^3,$$

as claimed. Next we establish the first result. Let $\boldsymbol{\delta}_0 = \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}$, and $t = \|\boldsymbol{\delta}\|$. Consider the composite function

$$\zeta(t) \doteq f(\exp_{\boldsymbol{q}}(t\boldsymbol{\delta}_0)) = f(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)),$$

and also

$$\begin{aligned}
\dot{\zeta}(t) &= \left\langle \nabla f\left(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)\right), -\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0\cos(t) \right\rangle \\
\ddot{\zeta}(t) &= \left\langle \nabla^2 f\left(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)\right)(-\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0\cos(t)), -\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0\cos(t) \right\rangle \\
&+ \quad \left\langle \nabla f\left(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)\right), -\boldsymbol{q}\cos(t) - \boldsymbol{\delta}_0\sin(t) \right\rangle .
\end{aligned}$$

In particular, this gives that

$$\begin{aligned}
\zeta(0) &= f(\boldsymbol{q}) \\
\dot{\zeta}(0) &= \langle \boldsymbol{\delta}_0, \nabla f(\boldsymbol{q}) \rangle \\
\ddot{\zeta}(0) &= \boldsymbol{\delta}_0^* \left(\nabla^2 f(\boldsymbol{q}) - \langle \nabla f(\boldsymbol{q}), \boldsymbol{q} \rangle \boldsymbol{I} \right) \boldsymbol{\delta}_0.
\end{aligned}$$

We next develop a bound on $\left|\ddot{\zeta}(t) - \ddot{\zeta}(0)\right|$. Using the triangle inequality, we can casually bound this difference as

$$\begin{aligned}
&\left|\ddot{\zeta}(t) - \ddot{\zeta}(0)\right| \\
\le\; &\left| \left\langle \nabla^2 f\left(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)\right)(-\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0\cos(t)), -\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0\cos(t) \right\rangle - \boldsymbol{\delta}_0^* \nabla^2 f(\boldsymbol{q}) \boldsymbol{\delta}_0 \right| \\
&\quad + \quad \left| \left\langle \nabla f\left(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)\right), -\boldsymbol{q}\cos(t) - \boldsymbol{\delta}_0\sin(t) \right\rangle + \langle \nabla f(\boldsymbol{q}), \boldsymbol{q} \rangle \right|
\end{aligned}$$

$$
\begin{aligned}
&\le \left|\left\langle \left[\nabla^2 f(\boldsymbol{q}\cos(t)+\boldsymbol{\delta}_0\sin(t)) - \nabla^2 f(\boldsymbol{q})\right](-\boldsymbol{q}\sin(t)+\boldsymbol{\delta}_0\cos(t)), -\boldsymbol{q}\sin(t)+\boldsymbol{\delta}_0\cos(t)\right\rangle\right| \\
&\qquad + \quad \left|\left\langle \nabla^2 f(\boldsymbol{q})\left(-\boldsymbol{q}\sin(t)+\boldsymbol{\delta}_0\cos(t)-\boldsymbol{\delta}_0\right), -\boldsymbol{q}\sin(t)+\boldsymbol{\delta}_0\cos(t)\right\rangle\right| \\
&\qquad + \quad \left|\left\langle \nabla^2 f(\boldsymbol{q})\boldsymbol{\delta}_0, -\boldsymbol{q}\sin(t)+\boldsymbol{\delta}_0\cos(t)-\boldsymbol{\delta}_0\right\rangle\right| \\
&\qquad + \quad \left|\langle \nabla f(\boldsymbol{q}\cos(t)+\boldsymbol{\delta}_0\sin(t)), -\boldsymbol{q}\cos(t)-\boldsymbol{\delta}_0\sin(t)\rangle + \langle \nabla f(\boldsymbol{q}\cos(t)+\boldsymbol{\delta}_0\sin(t)), \boldsymbol{q}\rangle\right| \\
&\qquad + \quad \left|\langle \nabla f(\boldsymbol{q}\cos(t)+\boldsymbol{\delta}_0\sin(t)), \boldsymbol{q}\rangle - \langle \nabla f(\boldsymbol{q}), \boldsymbol{q}\rangle\right| \\
&\le L_{\nabla^2}\left\|\boldsymbol{q}\cos(t)+\boldsymbol{\delta}_0\sin(t)-\boldsymbol{q}\right\| \\
&\qquad + M_{\nabla^2}\left\|-\boldsymbol{q}\sin(t)+\boldsymbol{\delta}_0\cos(t)-\boldsymbol{\delta}_0\right\| \\
&\qquad + M_{\nabla^2}\left\|-\boldsymbol{q}\sin(t)+\boldsymbol{\delta}_0\cos(t)-\boldsymbol{\delta}_0\right\| \\
&\qquad + M_{\nabla}\left\|-\boldsymbol{q}\cos(t)-\boldsymbol{\delta}_0\sin(t)+\boldsymbol{q}\right\| \\
&\qquad + L_{\nabla}\left\|\boldsymbol{q}\cos(t)+\boldsymbol{\delta}_0\sin(t)-\boldsymbol{q}\right\| \\
&= \left(L_{\nabla^2}+2M_{\nabla^2}+M_{\nabla}+L_{\nabla}\right)\sqrt{(1-\cos(t))^2+\sin^2(t)} \\
&= \eta_f\sqrt{2-2\cos t} \le \eta_f\sqrt{4\sin^2(t/2)} \le \eta_f t,
\end{aligned}
$$

where in the final line we have used the fact $1-\cos x = 2\sin^2(x/2)$ and that $\sin x \le x$ for $x \in [0,1]$, and M_{∇}, M_{∇^2}, L_{∇} and L_{∇^2} are the quantities defined in Lemma 5.3. By the integral form of Taylor's theorem in Lemma B.9 and the result above, we have

$$
\begin{aligned}
\left|f\left(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})\right) - \widehat{f}(\boldsymbol{q},\boldsymbol{\delta})\right| &= \left|\zeta(t) - \left(\zeta(0)+t\dot{\zeta}(0)+\tfrac{t^2}{2}\ddot{\zeta}(0)\right)\right| \\
&= \left|t^2\int_0^1 (1-s)\,\ddot{\zeta}(st)\;ds - \tfrac{t^2}{2}\ddot{\zeta}(0)\right| \\
&= t^2\left|\int_0^1 (1-s)\left[\ddot{\zeta}(st)-\ddot{\zeta}(0)\right]\;ds\right| \\
&\le t^2\int_0^1 (1-s)\,st\eta_f\;ds = \frac{\eta_f t^3}{6},
\end{aligned}
$$

with $t = \|\boldsymbol{\delta}\|$ we obtain the desired result. ■

10.3 Proof of Lemma 5.5

Proof By the integral form of Taylor's theorem in Lemma B.9, for any $t \in \left[0, \frac{3\Delta}{2\pi\sqrt{n}}\right]$, we have

$$
g\left(\boldsymbol{w} - t\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)
$$

$$
\begin{aligned}
&= g(\boldsymbol{w}) - t\int_0^1 \left\langle \nabla g\left(\boldsymbol{w} - st\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right), \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right\rangle\, ds \\
&= g\left(\boldsymbol{w}\right) - t\frac{\boldsymbol{w}^*\nabla g\left(\boldsymbol{w}\right)}{\|\boldsymbol{w}\|} + t\int_0^1 \left\langle \nabla g\left(\boldsymbol{w}\right) - \nabla g\left(\boldsymbol{w} - st\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right), \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right\rangle\, ds \\
&= g\left(\boldsymbol{w}\right) - t\frac{\boldsymbol{w}^*\nabla g\left(\boldsymbol{w}\right)}{\|\boldsymbol{w}\|} + t\int_0^1 \left(\left\langle \nabla g\left(\boldsymbol{w}\right), \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right\rangle - \left\langle \nabla g\left(\boldsymbol{w} - st\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right), \frac{\boldsymbol{w} - st\boldsymbol{w}/\left\|\boldsymbol{w}\right\|}{\left\|\boldsymbol{w} - st\boldsymbol{w}/\left\|\boldsymbol{w}\right\|\right\|}\right\rangle\right)\, ds \\
&\le g\left(\boldsymbol{w}\right) - t\frac{\boldsymbol{w}^*\nabla g\left(\boldsymbol{w}\right)}{\|\boldsymbol{w}\|} + \frac{L_g}{2}t^2 \le g\left(\boldsymbol{w}\right) - t\beta_g + \frac{L_g}{2}t^2.
\end{aligned}
$$

Minimizing this function over $t \in \left[0, \frac{3\Delta}{2\pi\sqrt{n}}\right]$, we obtain that there exists a $\boldsymbol{w}' \in \mathcal{B}\left(\boldsymbol{w}, \frac{3\Delta}{2\pi\sqrt{n}}\right)$ such that

$$
g(\boldsymbol{w}') \;\le\; g(\boldsymbol{w}) - \min\left\{\frac{\beta_g^2}{2L_g}, \frac{3\beta_g\Delta}{4\pi\sqrt{n}}\right\}.
$$

Given such a $\boldsymbol{w}' \in \mathcal{B}\left(\boldsymbol{w}, \frac{3\Delta}{2\pi\sqrt{n}}\right)$, there must exist some $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ such that $\boldsymbol{q}(\boldsymbol{w}') = \exp_{\boldsymbol{q}}(\boldsymbol{\delta})$. It remains to show that $\|\boldsymbol{\delta}\| \le \Delta$. By Lemma 9.7, we know that $\|\boldsymbol{q}(\boldsymbol{w}') - \boldsymbol{q}\left(\boldsymbol{w}\right)\| \le 2\sqrt{n}\left\|\boldsymbol{w}' - \boldsymbol{w}\right\| \le 3\Delta/\pi$. Hence,

$$
\left\|\exp_{\boldsymbol{q}}\left(\boldsymbol{\delta}\right) - \boldsymbol{q}\right\|^2 = \left\|\boldsymbol{q}\left(1 - \cos\left\|\boldsymbol{\delta}\right\|\right) + \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}\sin\left\|\boldsymbol{\delta}\right\|\right\|^2 = 2 - 2\cos\left\|\boldsymbol{\delta}\right\| = 4\sin^2\frac{\|\boldsymbol{\delta}\|}{2} \le \frac{9\Delta^2}{\pi^2},
$$

which means that $\sin\left(\|\boldsymbol{\delta}\|/2\right) \le 3\Delta/\left(2\pi\right)$. Because $\sin x \ge \frac{3}{\pi}x$ over $x \in [0, \pi/6]$, it implies that $\|\boldsymbol{\delta}\| \le \Delta$. Since $g(\boldsymbol{w}) = f(\boldsymbol{q}(\boldsymbol{w}))$, by summarizing all the results, we conclude that there exists a $\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\| \le \Delta$, such that

$$
f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) \le f(\boldsymbol{q}) - \min\left\{\frac{\beta_g^2}{2L_g}, \frac{3\beta_g\Delta}{4\pi\sqrt{n}}\right\},
$$

as claimed. ■

10.4 Proof of Lemma 5.6

Proof Let $\sigma = \operatorname{sign}\left(\boldsymbol{w}^*\nabla g(\boldsymbol{w})\right)$. For any $t \in \left[0, \frac{\Delta}{2\sqrt{n}}\right]$, by integral form of Taylor's theorem in Lemma B.9, we have

$$
\begin{aligned}
&g\left(\boldsymbol{w} - t\sigma\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) \\
&= g(\boldsymbol{w}) - t\sigma\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} + t^2\int_0^1 (1-s)\,\frac{\boldsymbol{w}^*\nabla^2 g\left(\boldsymbol{w} - st\sigma\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)\boldsymbol{w}}{\|\boldsymbol{w}\|^2}\, ds \\
&\le g(\boldsymbol{w}) + \frac{t^2}{2}\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w}}{\|\boldsymbol{w}\|^2} + t^2\int_0^1 \left[(1-s)\,\frac{\boldsymbol{w}^*\nabla^2 g\left(\boldsymbol{w} - st\sigma\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)\boldsymbol{w}}{\|\boldsymbol{w}\|^2} - (1-s)\,\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w}}{\|\boldsymbol{w}\|^2}\right]\, ds
\end{aligned}
$$

$$
\begin{aligned}
&= g(\boldsymbol{w}) + \frac{t^2}{2}\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w}}{\|\boldsymbol{w}\|^2} \\
&\quad + t^2\int_0^1 (1-s)\left[\frac{\left(\boldsymbol{w} - st\sigma\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)^*\nabla^2 g\left(\boldsymbol{w} - st\sigma\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)\left(\boldsymbol{w} - st\sigma\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)}{\left\|\boldsymbol{w} - st\sigma\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right\|^2} - \frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w}}{\|\boldsymbol{w}\|^2}\right] ds \\
&\le g(\boldsymbol{w}) - \frac{t^2}{2}\beta_{\frown} + t^2\int_0^1 (1-s)\, sL_{\frown}t\, ds \;\le\; g(\boldsymbol{w}) - \frac{t^2}{2}\beta_{\frown} + \frac{t^3}{6}L_{\frown}.
\end{aligned}
$$

Minimizing this function over $t \in \left[0, \frac{3\Delta}{2\pi\sqrt{n}}\right]$, we obtain

$$
t_\star = \min\left\{\frac{2\beta_{\frown}}{L_{\frown}}, \frac{3\Delta}{2\pi\sqrt{n}}\right\},
$$

and there exists a $\boldsymbol{w}' = \boldsymbol{w} - t_\star\sigma\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$ such that

$$
g\left(\boldsymbol{w} - t_\star\sigma\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) \;\le\; g(\boldsymbol{w}) - \min\left\{\frac{2\beta_{\frown}^3}{3L_{\frown}^2}, \frac{3\Delta^2\beta_{\frown}}{8\pi^2 n}\right\}.
$$

By arguments identical to those used in Lemma 5.5, there exists a tangent vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ such that $\boldsymbol{q}(\boldsymbol{w}') = \exp_{\boldsymbol{q}}(\boldsymbol{\delta})$ and $\|\boldsymbol{\delta}\| \le \Delta$. This completes the proof. ∎

10.5 Proof of Lemma 5.8

Proof For any $t \in \left[0, \frac{\Delta}{\left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|}\right]$, it holds that $\left\|t \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\| \le \Delta$, and the quadratic approximation

$$
\begin{aligned}
\widehat{f}\left(\boldsymbol{q}^{(k)}, -t \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right) &\le f\left(\boldsymbol{q}^{(k)}\right) - t\left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|^2 + \frac{M_H}{2}t^2\left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|^2 \\
&= f\left(\boldsymbol{q}^{(k)}\right) - t\left(1 - \frac{1}{2}M_H t\right)\left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|^2.
\end{aligned}
$$

Taking $t_0 = \min\left\{\frac{\Delta}{\left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|}, \frac{1}{M_H}\right\}$, we obtain

$$
\widehat{f}\left(\boldsymbol{q}^{(k)}, -t_0 \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right) \le f\left(\boldsymbol{q}^{(k)}\right) - \frac{1}{2}\min\left\{\frac{\Delta}{\left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|}, \frac{1}{M_H}\right\}\left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|^2. \tag{10.5.1}
$$

Now let $\boldsymbol{U}$ be an arbitrary orthonormal basis for $T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}$. Since the norm constraint is active, by the optimality condition in (5.3.5), we have

$$
\begin{aligned}
\Delta &\le \left\|\left[\boldsymbol{U}^* \operatorname{Hess} f\left(\boldsymbol{q}^{(k)}\right)\boldsymbol{U}\right]^{-1}\boldsymbol{U}^* \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\| \\
&\le \left\|\left[\boldsymbol{U}^* \operatorname{Hess} f\left(\boldsymbol{q}^{(k)}\right)\boldsymbol{U}\right]^{-1}\right\|\left\|\boldsymbol{U}^* \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\| \le \frac{\left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|}{m_H},
\end{aligned}
$$

which means that $\left\|\mathrm{grad}\, f\left(\boldsymbol{q}^{(k)}\right)\right\| \geq m_H \Delta$. Substituting this into (10.5.1), we obtain

$$\widehat{f}\left(\boldsymbol{q}^{(k)}, -t_0\ \mathrm{grad}\, f\left(\boldsymbol{q}^{(k)}\right)\right) \leq f\left(\boldsymbol{q}^{(k)}\right) - \frac{1}{2}\min\left\{m_H \Delta^2, \frac{m_H^2}{M_H}\Delta^2\right\} \leq f\left(\boldsymbol{q}^{(k)}\right) - \frac{m_H^2 \Delta^2}{2M_H}.$$

By the key comparison result established in proof of Lemma 5.4, we have

$$\begin{aligned} f\left(\exp_{\boldsymbol{q}^{(k)}}\left(-t_0\, \mathrm{grad}\, f\left(\boldsymbol{q}^{(k)}\right)\right)\right) &\leq \widehat{f}\left(\boldsymbol{q}^{(k)}, -t_0\ \mathrm{grad}\, f\left(\boldsymbol{q}^{(k)}\right)\right) + \frac{1}{6}\eta_f \Delta^3 \\ &\leq f\left(\boldsymbol{q}^{(k)}\right) - \frac{m_H^2 \Delta^2}{M_H} + \frac{1}{6}\eta_f \Delta^3. \end{aligned}$$

This completes the proof. ■

10.6 Proof of Lemma 5.9

It takes certain delicate work to prove Lemma 5.9. Basically to use discretization argument, the degree of continuity of the Hessian is needed. The tricky part is that for continuity, we need to compare the Hessian operators at different points, while these Hessian operators are only defined on the respective tangent planes. This is the place where parallel translation comes into play. The next two lemmas compute spectral bounds for the forward and inverse parallel translation operators.

Lemma 10.1 *For $\tau \in [0,1]$ and $\|\boldsymbol{\delta}\| \leq 1/2$, we have*

$$\left\|\mathcal{P}_\gamma^{\tau \leftarrow 0} - \boldsymbol{I}\right\| \leq \frac{5}{4}\tau \|\boldsymbol{\delta}\|, \quad (10.6.1)$$

$$\left\|\mathcal{P}_\gamma^{0 \leftarrow \tau} - \boldsymbol{I}\right\| \leq \frac{3}{2}\tau \|\boldsymbol{\delta}\|. \quad (10.6.2)$$

Proof By (5.3.6), we have

$$\begin{aligned} \left\|\mathcal{P}_\gamma^{\tau \leftarrow 0} - \boldsymbol{I}\right\| &= \left\|(\cos(\tau\|\boldsymbol{\delta}\|) - 1)\frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2} - \sin(\tau\|\boldsymbol{\delta}\|)\frac{\boldsymbol{q}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|}\right\| \\ &\leq 1 - \cos(\tau\|\boldsymbol{\delta}\|) + \sin(\tau\|\boldsymbol{\delta}\|) \\ &\leq 2\sin^2\left(\frac{\tau\|\boldsymbol{\delta}\|}{2}\right) + \sin(\tau\|\boldsymbol{\delta}\|) \leq \frac{1}{4}\tau\|\boldsymbol{\delta}\| + \tau\|\boldsymbol{\delta}\| \leq \frac{5}{4}\tau\|\boldsymbol{\delta}\|, \end{aligned}$$

where we have used the fact $\sin(t) \leq t$ and $1 - \cos x = 2\sin^2(x/2)$. Moreover, $\mathcal{P}_\gamma^{0 \leftarrow \tau}$ is in the form of $(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^*)^{-1}$ for some vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. By the Sherman-Morrison matrix inverse formula, i.e., $(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^*)^{-1} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{v}^*/(1 + \boldsymbol{v}^*\boldsymbol{u})$ (justified as $\left\|(\cos(\tau\|\boldsymbol{\delta}\|) - 1)\frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2} - \boldsymbol{q}\sin(\tau\|\boldsymbol{\delta}\|)\frac{\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|}\right\| \leq 5\tau\|\boldsymbol{\delta}\|/4 \leq 5/8 < 1$ as shown above), we have

$$\begin{aligned}
&\left\|\mathcal{P}_{\gamma}^{0\leftarrow\tau} - \boldsymbol{I}\right\| \\
=\; &\left\|\left(\cos(\tau\left\|\boldsymbol{\delta}\right\|) - 1\right)\frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\left\|\boldsymbol{\delta}\right\|^2} - \boldsymbol{q}\sin\left(\tau\left\|\boldsymbol{\delta}\right\|\right)\frac{\boldsymbol{\delta}^*}{\left\|\boldsymbol{\delta}\right\|}\right\| \frac{1}{1+\left(\cos\left(\tau\left\|\boldsymbol{\delta}\right\|\right)-1\right)} \qquad (\text{as } \boldsymbol{q}^*\boldsymbol{\delta} = 0) \\
\le\; &\frac{5}{4}\tau\left\|\boldsymbol{\delta}\right\|\frac{1}{\cos\left(\tau\left\|\boldsymbol{\delta}\right\|\right)} \le \frac{5}{4}\tau\left\|\boldsymbol{\delta}\right\|\frac{1}{\cos\left(1/2\right)} \le \frac{3}{2}\tau\left\|\boldsymbol{\delta}\right\|,
\end{aligned}$$

completing the proof. ■

The next lemma establish the “local-Lipschitz" property of the Riemannian Hessian.

> **Lemma 10.2** *Let $\gamma(t) = \exp_{\boldsymbol{q}}\left(t\boldsymbol{\delta}\right)$ denotes a geodesic curve on $\mathbb{S}^{n-1}$. Whenever $\left\|\boldsymbol{\delta}\right\| \le 1/2$ and $\tau \in [0,1]$,*
>
> $$\left\|\mathcal{P}_{\gamma}^{0\leftarrow\tau}\,\mathrm{Hess}\, f(\gamma(\tau))\mathcal{P}_{\gamma}^{\tau\leftarrow 0} - \mathrm{Hess}\, f(\boldsymbol{q})\right\| \le L_H \cdot \tau\left\|\boldsymbol{\delta}\right\|, \tag{10.6.3}$$
>
> *where $L_H = \frac{5}{2\mu^2}n^{3/2}\left\|\boldsymbol{X}\right\|_\infty^3 + \frac{9}{\mu}n\left\|\boldsymbol{X}\right\|_\infty^2 + 9\sqrt{n}\left\|\boldsymbol{X}\right\|_\infty$.*

Proof First of all, by (5.3.4) and using the fact that the operator norm of a projection operator is unitary bounded, we have

$$\begin{aligned}
&\left\|\mathrm{Hess}\, f(\gamma(\tau)) - \mathrm{Hess}\, f(\boldsymbol{q})\right\| \\
\le\; &\left\|\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}}\left[\nabla^2 f\left(\gamma\left(\tau\right)\right) - \nabla^2 f\left(\boldsymbol{q}\right) - \left(\left\langle\nabla f\left(\gamma\left(\tau\right)\right), \gamma\left(\tau\right)\right\rangle - \left\langle\nabla f\left(\boldsymbol{q}\right), \boldsymbol{q}\right\rangle\right)\boldsymbol{I}\right]\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}}\right\| \\
&\quad + \Big\|\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}}\left(\nabla^2 f\left(\boldsymbol{q}\right) - \left\langle\nabla f\left(\boldsymbol{q}\right), \boldsymbol{q}\right\rangle\boldsymbol{I}\right)\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}} \\
&\quad - \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}\left(\nabla^2 f\left(\boldsymbol{q}\right) - \left\langle\nabla f\left(\boldsymbol{q}\right), \boldsymbol{q}\right\rangle\boldsymbol{I}\right)\mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}\Big\| \\
\le\; &\left\|\nabla^2 f\left(\gamma\left(\tau\right)\right) - \nabla^2 f\left(\boldsymbol{q}\right)\right\| + \left|\left\langle\nabla f\left(\gamma\left(\tau\right)\right) - \nabla f\left(\boldsymbol{q}\right), \gamma\left(\tau\right)\right\rangle\right| + \left|\left\langle\nabla f\left(\boldsymbol{q}\right), \gamma\left(\tau\right) - \boldsymbol{q}\right\rangle\right| \\
&\quad + \left\|\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}} - \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}\right\|\left\|\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}} + \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}\right\|\left\|\nabla^2 f\left(\boldsymbol{q}\right) - \left\langle\nabla f\left(\boldsymbol{q}\right), \boldsymbol{q}\right\rangle\boldsymbol{I}\right\|.
\end{aligned}$$

By the estimates in Lemma 5.3, we obtain

$$\begin{aligned}
&\left\|\mathrm{Hess}\, f(\gamma(\tau)) - \mathrm{Hess}\, f(\boldsymbol{q})\right\| \\
\le\; &\frac{2}{\mu^2}n^{3/2}\left\|\boldsymbol{X}\right\|_\infty^3\left\|\gamma\left(\tau\right) - \boldsymbol{q}\right\| + \frac{n}{\mu}\left\|\boldsymbol{X}\right\|_\infty^2\left\|\gamma\left(\tau\right) - \boldsymbol{q}\right\| + \sqrt{n}\left\|\boldsymbol{X}\right\|_\infty\left\|\gamma\left(\tau\right) - \boldsymbol{q}\right\| \\
&\quad + 2\left\|\gamma\left(\tau\right)\gamma^*\left(\tau\right) - \boldsymbol{q}\boldsymbol{q}^*\right\|\left(\frac{n}{\mu}\left\|\boldsymbol{X}\right\|_\infty^2 + \sqrt{n}\left\|\boldsymbol{X}\right\|_\infty\right) \\
\le\; &\left(\frac{5}{2\mu^2}n^{3/2}\left\|\boldsymbol{X}\right\|_\infty^3 + \frac{25n}{4\mu}\left\|\boldsymbol{X}\right\|_\infty^2 + \frac{25}{4}\sqrt{n}\left\|\boldsymbol{X}\right\|_\infty\right)\tau\left\|\boldsymbol{\delta}\right\|,
\end{aligned} \tag{10.6.4}$$

where at the last line we have used the following estimates:

$$\begin{aligned} \|\gamma(\tau) - \boldsymbol{q}\| &= \left\| \boldsymbol{q}\left(\cos\left(\tau\left\|\boldsymbol{\delta}\right\|\right) - 1\right) + \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}\sin\left(\tau\left\|\boldsymbol{\delta}\right\|\right)\right\| \le \frac{5}{4}\tau\left\|\boldsymbol{\delta}\right\|, \ \text{(Proof of Lemma 10.1)} \\ \|\gamma(\tau)\gamma^*(\tau) - \boldsymbol{q}\boldsymbol{q}^*\| &\le \left\| \left(\frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2} - \boldsymbol{q}\boldsymbol{q}^*\right)\sin^2\left(\tau\left\|\boldsymbol{\delta}\right\|\right)\right\| + 2\sin\left(\tau\left\|\boldsymbol{\delta}\right\|\right)\cos\left(\tau\left\|\boldsymbol{\delta}\right\|\right) \\ &\le \sin^2\left(\tau\left\|\boldsymbol{\delta}\right\|\right) + \sin\left(2\tau\left\|\boldsymbol{\delta}\right\|\right) \le \frac{5}{2}\tau\left\|\boldsymbol{\delta}\right\|. \end{aligned}$$

Therefore, by Lemma 10.1, we obtain

$$\begin{aligned} &\left\|\mathcal{P}_\gamma^{0\leftarrow\tau}\operatorname{Hess} f(\gamma(\tau))\mathcal{P}_\gamma^{\tau\leftarrow 0} - \operatorname{Hess} f(\boldsymbol{q})\right\| \\ \le\ &\left\|\mathcal{P}_\gamma^{0\leftarrow\tau}\operatorname{Hess} f(\gamma(\tau))\mathcal{P}_\gamma^{\tau\leftarrow 0} - \operatorname{Hess} f(\gamma(\tau))\mathcal{P}_\gamma^{\tau\leftarrow 0}\right\| + \left\|\operatorname{Hess} f(\gamma(\tau))\mathcal{P}_\gamma^{\tau\leftarrow 0} - \operatorname{Hess} f(\gamma(\tau))\right\| \\ &\quad + \left\|\operatorname{Hess} f(\gamma(\tau)) - \operatorname{Hess} f(\boldsymbol{q})\right\| \\ \le\ &\left\|\mathcal{P}_\gamma^{0\leftarrow\tau} - \boldsymbol{I}\right\|\left\|\operatorname{Hess} f(\gamma(\tau))\right\| + \left\|\mathcal{P}_\gamma^{\tau\leftarrow 0} - \boldsymbol{I}\right\|\left\|\operatorname{Hess} f(\gamma(t))\right\| + \left\|\operatorname{Hess} f(\gamma(t)) - \operatorname{Hess} f(\boldsymbol{q})\right\| \\ \le\ &\frac{11}{4}\tau\left\|\boldsymbol{\delta}\right\|\left\|\nabla^2 f\left(\gamma\left(\tau\right)\right) - \left\langle \nabla f\left(\gamma\left(\tau\right)\right), \gamma\left(t\right)\right\rangle \boldsymbol{I}\right\| + \left\|\operatorname{Hess} f(\gamma(\tau)) - \operatorname{Hess} f(\boldsymbol{q})\right\|. \end{aligned}$$

By Lemma 5.3 and substituting the estimate in (10.6.4), we obtain the claimed result. ■

Proof [of Lemma 5.9] For any given $\boldsymbol{q}$ with $\|\boldsymbol{w}(\boldsymbol{q})\| \le \mu/(4\sqrt{2})$, assume $\boldsymbol{U}$ is an orthonormal basis for its tangent space $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$. We could compare $\boldsymbol{U}^*\operatorname{Hess} f(\boldsymbol{q})\boldsymbol{U}$ with $\nabla^2_{\boldsymbol{w}} g(\boldsymbol{w})$, and build on the known results for the latter. Instead, we present a direct proof here that yields tighter results as stated in the lemma. Again we first work with the "canonical" section in the vicinity of $\boldsymbol{e}_n$ with the "canonical" reparametrization $\boldsymbol{q}(\boldsymbol{w}) = [\boldsymbol{w}; \sqrt{1 - \|\boldsymbol{w}\|^2}]$.

By definition of the Riemannian Hessian in (5.3.4), expressions of $\nabla^2 f$ and ∇f in (5.3.1) and (5.3.2), and exchange of differential and expectation operators (justified similarly as in Section 9.1.3), we obtain

$$\begin{aligned} \boldsymbol{U}^*\operatorname{Hess}\mathbb{E}\left[f(\boldsymbol{q})\right]\boldsymbol{U} &= \mathbb{E}\left[\boldsymbol{U}^*\operatorname{Hess} f(\boldsymbol{q})\boldsymbol{U}\right] \\ &= \mathbb{E}\left[\boldsymbol{U}^*\nabla^2 f(\boldsymbol{q})\boldsymbol{U} - \left\langle \boldsymbol{q}, \nabla f(\boldsymbol{q})\right\rangle \boldsymbol{I}_{n-1}\right] \\ &= \boldsymbol{U}^*\mathbb{E}\left[\frac{1}{\mu}\left\{1 - \tanh^2\left(\frac{\boldsymbol{q}^*\boldsymbol{x}}{\mu}\right)\right\}\boldsymbol{x}\boldsymbol{x}^*\right]\boldsymbol{U} - \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\boldsymbol{x}}{\mu}\right)\boldsymbol{q}^*\boldsymbol{x}\right]\boldsymbol{I}_{n-1}. \end{aligned}$$

We have

$$\boldsymbol{U}^*\mathbb{E}\left[\frac{1}{\mu}\left\{1 - \tanh^2\left(\frac{\boldsymbol{q}^*\boldsymbol{x}}{\mu}\right)\right\}\boldsymbol{x}\boldsymbol{x}^*\right]\boldsymbol{U} \succeq \frac{1-\theta}{\mu}\boldsymbol{U}^*\mathbb{E}\left[\left\{1 - \tanh^2\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}}{\mu}\right)\right\}\begin{bmatrix}\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^* & \boldsymbol{0} \\ \boldsymbol{0}^* & 0\end{bmatrix}\right]\boldsymbol{U}.$$

Now consider any vector $\boldsymbol{z} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ such that $\boldsymbol{z} = \boldsymbol{U}\boldsymbol{v}$ for some $\boldsymbol{v} \in \mathbb{R}^{n-1}$ and $\|\boldsymbol{z}\| = 1$. Then

$$\boldsymbol{z}^*\mathbb{E}\left[\left\{1 - \tanh^2\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}}{\mu}\right)\right\}\begin{bmatrix}\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^* & \boldsymbol{0} \\ \boldsymbol{0}^* & 0\end{bmatrix}\right]\boldsymbol{z} \geq \frac{\theta}{\sqrt{2\pi}}(2 - 3\sqrt{2}/4)\|\overline{\boldsymbol{z}}\|^2$$

by proof of Proposition 4.7, where $\overline{\boldsymbol{z}} \in \mathbb{R}^{n-1}$ as above is the first $n-1$ coordinates of $\boldsymbol{z}$. Now we know that $\langle \boldsymbol{q}, \boldsymbol{z}\rangle = 0$, or

$$\boldsymbol{w}^*\overline{\boldsymbol{z}} + q_n z_n = 0 \Longrightarrow \frac{\|\overline{\boldsymbol{z}}\|}{|z_n|} = \frac{q_n}{\|\boldsymbol{w}\|} = \frac{\sqrt{1 - \|\boldsymbol{w}\|^2}}{\|\boldsymbol{w}\|} \geq 50,$$

where we have used $\|\boldsymbol{w}\| \leq \mu/(4\sqrt{2})$ and $\mu \leq 1/10$ to obtain the last lower bound. Combining the above with the fact that $\|\boldsymbol{z}\| = 1$, we obtain

$$\boldsymbol{U}^*\mathbb{E}\left[\frac{1}{\mu}\left\{1 - \tanh^2\left(\frac{\boldsymbol{q}^*\boldsymbol{x}}{\mu}\right)\right\}\boldsymbol{x}\boldsymbol{x}^*\right]\boldsymbol{U} \succeq \frac{99}{100}\frac{1-\theta}{\mu}\frac{\theta}{\sqrt{2\pi}}(2 - 3\sqrt{2}/4)\boldsymbol{I}_{n-1} \tag{10.6.5}$$

$$\succeq \frac{99}{200\sqrt{2\pi}}(2 - 3\sqrt{2}/4)\frac{\theta}{\mu}\boldsymbol{I}_{n-1}, \tag{10.6.6}$$

where we have simplified the expression using $\theta \leq 1/2$. To bound the second term,

$$\begin{aligned}
&\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\boldsymbol{x}_k}{\mu}\right)\boldsymbol{q}^*\boldsymbol{x}_k\right] \\
&= \mathbb{E}_{\mathcal{I}}\left[\mathbb{E}_{Z\sim\mathcal{N}(0,\|\boldsymbol{q}_{\mathcal{I}}\|^2)}\left[\tanh(Z/\mu)Z\right]\right] \\
&= \frac{1}{\mu}\mathbb{E}_{\mathcal{I}}\left[\|\boldsymbol{q}_{\mathcal{I}}\|^2\mathbb{E}_{Z\sim\mathcal{N}(0,\|\boldsymbol{q}_{\mathcal{I}}\|^2)}\left[1 - \tanh^2(Z/\mu)\right]\right] \quad \text{(by Lemma B.10)} \\
&\leq \frac{1}{\mu}\mathbb{E}_{\mathcal{I}}\left[\mathbb{E}_{Z\sim\mathcal{N}(0,\|\boldsymbol{q}_{\mathcal{I}}\|^2)}\left[1 - \tanh^2(Z/\mu)\right]\right].
\end{aligned}$$

Now we have the following estimate:

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{N}(0,\|\boldsymbol{w}_{\mathcal{J}}\|^2+q_n^2)}\left[1 - \tanh^2(Z/\mu)\right] \\
&= 2\mathbb{E}_{Z\sim\mathcal{N}(0,\|\boldsymbol{w}_{\mathcal{J}}\|^2+q_n^2)}\left[\left(1 - \tanh^2(Z/\mu)\right)\mathbb{1}_{Z>0}\right] \\
&\leq 8\mathbb{E}_{Z\sim\mathcal{N}(0,\|\boldsymbol{w}_{\mathcal{J}}\|^2+q_n^2)}\left[\exp(-2Z/\mu)\mathbb{1}_{Z>0}\right] \\
&= 8\exp\left(\frac{2\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2 + 2q_n^2}{\mu^2}\right)\Phi^c\left(\frac{2\sqrt{\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2 + q_n^2}}{\mu}\right) \quad \text{(by Lemma B.10)} \\
&\leq \frac{4}{\sqrt{2\pi}}\frac{\mu}{\sqrt{\left\|\boldsymbol{w}_{\mathcal{J}}\right\|^2 + q_n^2}},
\end{aligned}$$

where at the last inequality we have applied Gaussian tail upper bound of Type II in Lemma B.5. Since

$\|\boldsymbol{w}_{\mathcal{J}}\|^2 + q_n^2 \geq q_n^2 = 1 - \|\boldsymbol{w}\|^2 \geq 1 - \mu^2/32 \geq 31/32$ for $\|\boldsymbol{w}\| \leq \mu/(4\sqrt{2})$ and $\mu \leq 1$, we obtain

$$\mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{w}_{\mathcal{J}}\|^2+q_n^2\right)}\left[1-\tanh^2(Z/\mu)\right] \leq \frac{4}{\sqrt{2\pi}}\frac{\mu}{\sqrt{31/32}} \leq \frac{4}{\sqrt{2\pi}}\mu. \tag{10.6.7}$$

Collecting the above estimates, we obtain

$$\boldsymbol{U}^*\,\mathrm{Hess}\,\mathbb{E}\left[f(\boldsymbol{q})\right]\boldsymbol{U} \succeq \frac{99}{200\sqrt{2\pi}}(2-3\sqrt{2}/4)\frac{\theta}{\mu}\boldsymbol{I}_{n-1} - \frac{1}{\mu}\frac{4}{\sqrt{2\pi}}\mu\boldsymbol{I}_{n-1} \succeq \frac{1}{4\sqrt{2\pi}}\frac{\theta}{\mu}\boldsymbol{I}_{n-1}, \tag{10.6.8}$$

where we have used the fact $\mu \leq \theta/10$ to obtain the final lower bound.

Next we perform concentration analysis. For any $\boldsymbol{q}$, we can write

$$\boldsymbol{U}^*\nabla^2 f(\boldsymbol{q})\boldsymbol{U} = \frac{1}{p}\sum_{k=1}^{p}\boldsymbol{W}_k, \quad \text{with } \boldsymbol{W}_k \doteq \frac{1}{\mu}\left[1-\tanh^2\left(\frac{\boldsymbol{q}^*\boldsymbol{x}_k}{\mu}\right)\right]\boldsymbol{U}^*\boldsymbol{x}_k\boldsymbol{x}_k^*\boldsymbol{U}.$$

For any integer $m \geq 2$, we have

$$\boldsymbol{0} \preceq \mathbb{E}\left[\boldsymbol{W}_k^m\right] \preceq \frac{1}{\mu^m}\mathbb{E}\left[(\boldsymbol{U}^*\boldsymbol{x}_k\boldsymbol{x}_k^*\boldsymbol{U})^m\right] \preceq \frac{1}{\mu^m}\mathbb{E}\left[\|\boldsymbol{x}_k\boldsymbol{x}_k^*\|^m\right]\boldsymbol{I} = \frac{1}{\mu^m}\mathbb{E}\left[\|\boldsymbol{x}_k\|^{2m}\right]\boldsymbol{I} \preceq \frac{1}{\mu^m}\mathbb{E}_{Z\sim\xi^2(n)}\left[Z^m\right]\boldsymbol{I},$$

where we have used Lemma 9.4 to obtain the last inequality. By Lemma B.7, we obtain

$$\boldsymbol{0} \preceq \mathbb{E}\left[\boldsymbol{W}_k^m\right] \preceq \frac{1}{\mu^m}\frac{m!}{2}(2n)^m\boldsymbol{I} \preceq \frac{m!}{2}\left(\frac{2n}{\mu}\right)^m\boldsymbol{I}.$$

Taking $R_{\boldsymbol{W}} = 2n/\mu$, and $\sigma_{\boldsymbol{W}}^2 = 4n^2/\mu^2 \geq \mathbb{E}\left[\boldsymbol{W}_k^2\right]$, by Lemma A.2, we obtain

$$\mathbb{P}\left[\left\|\frac{1}{p}\sum_{k=1}^{p}\boldsymbol{W}_k - \frac{1}{p}\sum_{k=1}^{p}\mathbb{E}\left[\boldsymbol{W}_k\right]\right\| > t/2\right] \leq 2n\exp\left(-\frac{p\mu^2t^2}{32n^2+8nt}\right) \tag{10.6.9}$$

for any $t > 0$. Similarly, we write

$$\langle\nabla f(\boldsymbol{q}),\boldsymbol{q}\rangle = \frac{1}{p}\sum_{k=1}^{p}Z_k, \quad \text{with } Z_k \doteq \tanh\left(\frac{\boldsymbol{q}^*\boldsymbol{x}_k}{\mu}\right)\boldsymbol{q}^*\boldsymbol{x}_k.$$

For any integer $m \geq 2$, we have

$$\mathbb{E}\left[|Z_k|^m\right] \leq \mathbb{E}\left[|\boldsymbol{q}^*\boldsymbol{x}_k|^m\right] \leq \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[|Z|^m\right] \leq \frac{m!}{2},$$

where at the first inequality we used the fact $|\tanh(\cdot)| \leq 1$, at the second we invoked Lemma 9.4, and at the third we invoked Lemma B.6. Taking $R_Z = \sigma_Z^2 = 1$, by Lemma A.1, we obtain

$$\mathbb{P}\left[\left|\frac{1}{p}\sum_{k=1}^{p}Z_k - \frac{1}{p}\sum_{k=1}^{p}\mathbb{E}\left[Z_k\right]\right| > t/2\right] \leq 2\exp\left(-pt^2/16\right) \tag{10.6.10}$$

for any $t > 0$. Gathering (10.6.9) and (10.6.10), we obtain that for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\|\boldsymbol{U}^* \operatorname{Hess} \mathbb{E}\left[f(\boldsymbol{q})\right] \boldsymbol{U} - \boldsymbol{U}^* \operatorname{Hess} f(\boldsymbol{q}) \boldsymbol{U}\| > t\right] \\
\le\ & \mathbb{P}\left[\left\|\boldsymbol{U}^* \nabla^2 f(\boldsymbol{q}) \boldsymbol{U} - \nabla^2 \mathbb{E}\left[f(\boldsymbol{q})\right]\right\| > t/2\right] + \mathbb{P}\left[\left|\langle \nabla f(\boldsymbol{q}), \boldsymbol{q}\rangle - \langle \nabla \mathbb{E}\left[f(\boldsymbol{q})\right], \boldsymbol{q}\rangle\right| > t/2\right] \\
\le\ & 2n \exp\left(-\frac{p\mu^2 t^2}{32n^2 + 8nt}\right) + 2\exp\left(-\frac{pt^2}{16}\right) \le 4n \exp\left(-\frac{p\mu^2 t^2}{32n^2 + 8nt}\right). \qquad (10.6.11)
\end{aligned}$$

Now we are ready to pull above results together for a discretization argument. For any $\varepsilon \in (0, \mu/(4\sqrt{2}))$, there is an ε-net N_ε of size at most $(3\mu/(4\sqrt{2}\varepsilon))^n$ that covers the region $\left\{\boldsymbol{q} : \|\boldsymbol{w}(\boldsymbol{q})\| \le \mu/(4\sqrt{2})\right\}$. By Lemma 10.2, the function $\operatorname{Hess} f(\boldsymbol{q})$ is locally Lipschitz within each normal ball of radius

$$\left\|\boldsymbol{q} - \exp_{\boldsymbol{q}}(1/2)\right\| = \sqrt{2 - 2\cos(1/2)} \ge 1/\sqrt{5}$$

with Lipschitz constant L_H (as defined in Lemma 10.2). Note that $\varepsilon < \mu/(4\sqrt{2}) < 1/(4\sqrt{2}) < 1/\sqrt{5}$ for $\mu < 1$, so any choice of $\varepsilon \in (0, \mu/(4\sqrt{2}))$ makes the Lipschitz constant L_H valid within each ε-ball centered around one element of the ε-net. Let

$$\mathcal{E}_\infty \doteq \left\{1 \le \|\boldsymbol{X}_0\|_\infty \le 4\sqrt{\log(np)}\right\}.$$

From Lemma 9.11, $\mathbb{P}\left[\mathcal{E}_\infty^c\right] \le \theta\,(np)^{-7} + \exp\left(-0.3\theta np\right)$. By Lemma 10.2, with at least the same probability,

$$L_H \le C_1 \frac{n^{3/2}}{\mu^2} \log^{3/2}(np).$$

Set $\varepsilon = \frac{\theta}{12\sqrt{2\pi}\mu L_H} < \mu/(4\sqrt{2})$, so

$$\# N_\varepsilon \le \exp\left(n \log \frac{C_2 n^{3/2} \log^{3/2}(np)}{\theta}\right).$$

Let $\mathcal{E}_H$ denote the event that

$$\mathcal{E}_H \doteq \left\{\max_{\boldsymbol{q} \in N_\varepsilon} \|\boldsymbol{U}^* \operatorname{Hess} \mathbb{E}\left[f(\boldsymbol{q})\right] \boldsymbol{U} - \boldsymbol{U}^* \operatorname{Hess} f(\boldsymbol{q}) \boldsymbol{U}\| \le \frac{\theta}{12\sqrt{2\pi}\mu}\right\}.$$

On $\mathcal{E}_\infty \cap \mathcal{E}_H$,

$$\sup_{\boldsymbol{q} : \|\boldsymbol{w}(\boldsymbol{q})\| \le \mu/(4\sqrt{2})} \|\boldsymbol{U}^* \operatorname{Hess} \mathbb{E}\left[f(\boldsymbol{q})\right] \boldsymbol{U} - \boldsymbol{U}^* \operatorname{Hess} f(\boldsymbol{q}) \boldsymbol{U}\| \le \frac{\theta}{6\sqrt{2\pi}\mu}.$$

So on $\mathcal{E}_\infty \cap \mathcal{E}_H$, we have

$$\boldsymbol{U}^* \operatorname{Hess} f(\boldsymbol{q}) \boldsymbol{U} \succeq c_\sharp \frac{\theta}{\mu} \qquad (10.6.12)$$

for any $c_\sharp \le 1/(12\sqrt{2\pi})$. Setting $t = \frac{\theta}{12\sqrt{2\pi}\mu}$ in (10.6.11), we obtain that for any fixed $\boldsymbol{q}$ in this region,

$$\mathbb{P}\left[\left\|\boldsymbol{U}^* \operatorname{Hess} \mathbb{E}\left[f(\boldsymbol{q})\right] \boldsymbol{U} - \boldsymbol{U}^* \operatorname{Hess} f(\boldsymbol{q}) \boldsymbol{U}\right\| > t\right] \le 4n \exp\left(-\frac{p\theta^2}{c_3 n^2 + c_4 n\theta/\mu}\right).$$

Taking a union bound, we obtain that

$$\mathbb{P}\left[\mathcal{E}_H^c\right] \le 4n \exp\left(-\frac{p\theta^2}{c_3 n^2 + c_4 n\theta/\mu} + C_5 n \log n + C_6 n \log\log p\right).$$

It is enough to make $p \ge C_7 n^3 \log(n/(\mu\theta))/(\mu\theta^2)$ to make the failure probability small, completing the proof.

■

10.7 Proof of Lemma 5.11

Proof For a given $\boldsymbol{q}$, consider the vector $\boldsymbol{r} \doteq \boldsymbol{q} - \boldsymbol{e}_n/q_n$. It is easy to verify that $\langle \boldsymbol{q}, \boldsymbol{r}\rangle = 0$, and hence $\boldsymbol{r} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$. Now, by (5.3.1) and (5.3.3), we have

$$\begin{aligned}
\langle \operatorname{grad} f(\boldsymbol{q}), \boldsymbol{r}\rangle &= \langle (\boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^*) \nabla f(\boldsymbol{q}), \boldsymbol{q} - \boldsymbol{e}_n/q_n\rangle \\
&= \langle (\boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^*) \nabla f(\boldsymbol{q}), -\boldsymbol{e}_n/q_n\rangle \\
&= \frac{1}{p}\sum_{k=1}^{p} \left\langle (\boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^*) \tanh\left(\frac{\boldsymbol{q}^*\boldsymbol{x}_k}{\mu}\right) \boldsymbol{x}_k, -\boldsymbol{e}_n/q_n \right\rangle \\
&= \frac{1}{p}\sum_{k=1}^{p} \tanh\left(\frac{\boldsymbol{q}^*\boldsymbol{x}_k}{\mu}\right)\left(-\frac{x_k(n)}{q_n} + \boldsymbol{q}^*\boldsymbol{x}_k\right) \\
&= \frac{1}{p}\sum_{k=1}^{p} \tanh\left(\frac{\boldsymbol{q}^*\boldsymbol{x}_k}{\mu}\right)\left(\boldsymbol{w}^*(\boldsymbol{q})\overline{\boldsymbol{x}}_k - \frac{x_k(n)}{q_n}\left\|\boldsymbol{w}(\boldsymbol{q})\right\|^2\right) \\
&= \boldsymbol{w}^*(\boldsymbol{q}) \nabla g(\boldsymbol{w}),
\end{aligned}$$

where to get the last line we have used (9.0.1). Thus,

$$\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} = \frac{\langle \operatorname{grad} f(\boldsymbol{q}), \boldsymbol{r}\rangle}{\|\boldsymbol{w}\|} \le \|\operatorname{grad} f(\boldsymbol{q})\| \frac{\|\boldsymbol{r}\|}{\|\boldsymbol{w}\|},$$

where

$$\frac{\|\boldsymbol{r}\|^2}{\|\boldsymbol{w}\|^2} = \frac{\|\boldsymbol{w}\|^2 + \left(q_n - \frac{1}{q_n}\right)^2}{\|\boldsymbol{w}\|^2} = \frac{\|\boldsymbol{w}\|^2 + \|\boldsymbol{w}\|^4/q_n^2}{\|\boldsymbol{w}\|^2} = \frac{1}{q_n^2} = \frac{1}{1 - \|\boldsymbol{w}\|^2} \le \frac{1}{1 - \frac{1}{2000}} = \frac{2000}{1999},$$

where we have invoked our assumption that $\|\boldsymbol{w}\| \le \frac{1}{20\sqrt{5}}$. Therefore we obtain

$$\|\mathrm{grad}\, f(\boldsymbol{q})\| \ge \frac{\|\boldsymbol{w}\|}{\|\boldsymbol{r}\|} \frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} \ge \sqrt{\frac{1999}{2000}} \frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} \ge \frac{9}{10} \frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|},$$

completing the proof. ■

10.8 Proof of Lemma 5.12

Proof of Lemma 5.12 combines the local Lipschitz property of $\mathrm{Hess}\, f(\boldsymbol{q})$ in Lemma 10.2, and the Taylor's theorem (manifold version, Lemma 7.4.7 of [AMS09]).

Proof Let $\gamma(t)$ be the unique geodesic that satisfies $\gamma(0) = \boldsymbol{q}^{(k)}$, $\gamma(1) = \boldsymbol{q}^{(k+1)}$, and its directional derivative $\dot{\gamma}(0) = \boldsymbol{\delta}_\star$. Since the parallel translation defined by the Riemannian connection is an isometry, then $\left\|\mathrm{grad}\, f(\boldsymbol{q}^{(k+1)})\right\| = \left\|\mathcal{P}_\gamma^{0\leftarrow 1}\, \mathrm{grad}\, f(\boldsymbol{q}^{(k+1)})\right\|$. Moreover, since $\|\boldsymbol{\delta}_\star\| \le \Delta$, the unconstrained optimality condition in (5.3.5) implies that $\mathrm{grad}\, f(\boldsymbol{q}^{(k)}) + \mathrm{Hess}\, f(\boldsymbol{q}^{(k)})\boldsymbol{\delta}_\star = \boldsymbol{0}_{\boldsymbol{q}^{(k)}}$. Thus, by using Taylor's theorem in [AMS09], we have

$$\begin{aligned}
\left\|\mathrm{grad}\, f(\boldsymbol{q}^{(k+1)})\right\| &= \left\|\mathcal{P}_\gamma^{0\leftarrow 1}\, \mathrm{grad}\, f\left(\boldsymbol{q}^{(k+1)}\right) - \mathrm{grad}\, f\left(\boldsymbol{q}^{(k)}\right) - \mathrm{Hess}\, f\left(\boldsymbol{q}^{(k)}\right)\boldsymbol{\delta}_\star\right\| \\
&= \left\|\int_0^1 \left[\mathcal{P}_\gamma^{0\leftarrow t}\, \mathrm{Hess}\, f(\gamma(t))\,[\dot{\gamma}(t)] - \mathrm{Hess}\, f\left(\boldsymbol{q}^{(k)}\right)\boldsymbol{\delta}_\star\right] dt\right\| \quad \text{(Taylor's theorem)} \\
&= \left\|\int_0^1 \left(\mathcal{P}_\gamma^{0\leftarrow t}\, \mathrm{Hess}\, f(\gamma(t))\, \mathcal{P}_\gamma^{t\leftarrow 0}\boldsymbol{\delta}_\star - \mathrm{Hess}\, f\left(\boldsymbol{q}^{(k)}\right)\boldsymbol{\delta}_\star\right) dt\right\| \\
&\le \|\boldsymbol{\delta}_\star\| \int_0^1 \left\|\mathcal{P}_\gamma^{0\leftarrow t}\, \mathrm{Hess}\, f(\gamma(t))\, \mathcal{P}_\gamma^{t\leftarrow 0} - \mathrm{Hess}\, f\left(\boldsymbol{q}^{(k)}\right)\right\| dt.
\end{aligned}$$

From the Lipschitz bound in Lemma 10.2 and the optimality condition in (5.3.5), we obtain

$$\left\|\mathrm{grad}\, f\left(\boldsymbol{q}^{(k+1)}\right)\right\| \le \frac{1}{2}\|\boldsymbol{\delta}_\star\|^2 L_H = \frac{L_H}{2m_H^2}\left\|\mathrm{grad}\, f\left(\boldsymbol{q}^{(k)}\right)\right\|^2.$$

This completes the proof. ■

10.9 Proof of Lemma 5.14

Proof By invoking Taylor's theorem in [AMS09], we have

$$\mathcal{P}_\gamma^{0\leftarrow\tau}\, \mathrm{grad}\, f(\gamma(\tau)) = \int_0^\tau \mathcal{P}_\gamma^{0\leftarrow t}\, \mathrm{Hess}\, f(\gamma(t))\,[\dot{\gamma}(t)]\, dt.$$

Hence, we have

$$\begin{aligned}\left\langle \mathcal{P}_{\gamma}^{0\leftarrow\tau}\operatorname{grad} f\left(\gamma\left(\tau\right)\right), \boldsymbol{\delta}\right\rangle &= \int_0^{\tau} \left\langle \mathcal{P}_{\gamma}^{0\leftarrow t}\operatorname{Hess} f\left(\gamma\left(t\right)\right)[\dot{\gamma}\left(t\right)], \boldsymbol{\delta}\right\rangle\, dt \\ &= \int_0^{\tau} \left\langle \mathcal{P}_{\gamma}^{0\leftarrow t}\operatorname{Hess} f\left(\gamma\left(t\right)\right)[\dot{\gamma}\left(t\right)], \mathcal{P}_{\gamma}^{0\leftarrow t}\dot{\gamma}\left(t\right)\right\rangle\, dt \\ &= \int_0^{\tau} \left\langle \operatorname{Hess} f\left(\gamma\left(t\right)\right)[\dot{\gamma}\left(t\right)], \dot{\gamma}\left(t\right)\right\rangle\, dt \\ &\geq m_H \int_0^{\tau} \left\|\dot{\gamma}\left(t\right)\right\|^2\, dt \geq m_H \tau \left\|\boldsymbol{\delta}\right\|^2,\end{aligned}$$

where we have used the fact that the parallel transport $\mathcal{P}_{\gamma}^{0\leftarrow t}$ defined by the Riemannian connection is an isometry. On the other hand, we have

$$\left\langle \mathcal{P}_{\gamma}^{0\leftarrow\tau}\operatorname{grad} f\left(\gamma\left(\tau\right)\right), \boldsymbol{\delta}\right\rangle \leq \left\|\mathcal{P}_{\gamma}^{0\leftarrow\tau}\operatorname{grad} f\left(\gamma\left(\tau\right)\right)\right\| \left\|\boldsymbol{\delta}\right\| = \left\|\operatorname{grad} f\left(\gamma\left(\tau\right)\right)\right\| \left\|\boldsymbol{\delta}\right\|,$$

where again used the isometry property of the operator $\mathcal{P}_{\gamma}^{0\leftarrow\tau}$. Combining the two bounds above, we obtain

$$\left\|\operatorname{grad} f\left(\gamma\left(\tau\right)\right)\right\| \left\|\boldsymbol{\delta}\right\| \geq m_H \tau \left\|\boldsymbol{\delta}\right\|^2,$$

which implies the claimed result. ■

Chapter 11

Proofs of Technical Results for the Whole Recovery Pipeline

> Mathematics is not a deductive science – that's a cliche. When you try to prove a theorem, you don't just list the hypotheses, and then start to reason. What you do is trial and error, experimentation, guesswork.
>
> Paul Halmos

We need one technical lemma to prove Lemma 6.2 and the relevant lemma for complete dictionaries.

Lemma 11.1 *There exists a positive constant C, such that for all integer $n_1 \in \mathbb{N}$, $\theta \in (0, 1/3)$, and $n_2 \in \mathbb{N}$ with $n_2 \geq C n_1 \log\left(n_1/\theta\right)/\theta^2$, any random matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2} \sim_{i.i.d.} \mathrm{BG}(\theta)$ obeys the following. For any fixed index set $\mathcal{I} \subset [n_2]$ with $|\mathcal{I}| \leq \frac{9}{8}\theta n_2$, it holds that*

$$\left\|\boldsymbol{v}^* \boldsymbol{M}_{\mathcal{I}^c}\right\|_1 - \left\|\boldsymbol{v}^* \boldsymbol{M}_{\mathcal{I}}\right\|_1 \geq \frac{n_2}{6}\sqrt{\frac{2}{\pi}}\theta \left\|\boldsymbol{v}\right\| \quad \textit{for all } \boldsymbol{v} \in \mathbb{R}^{n_1},$$

with probability at least $1 - n_2^{-10} - \theta\left(n_1 n_2\right)^{-7} - \exp\left(-0.3\theta n_1 n_2\right)$.

Proof By homogeneity, it is sufficient to consider all $\boldsymbol{v} \in \mathbb{S}^{n_1 - 1}$. For any $i \in [n_2]$, let $\boldsymbol{m}_i \in \mathbb{R}^{n_1}$ be a column of $\boldsymbol{M}$. For a fixed $\boldsymbol{v}$ such that $\|\boldsymbol{v}\| = 1$, we have

$$T\left(\boldsymbol{v}\right) \doteq \left\|\boldsymbol{v}^* \boldsymbol{M}_{\mathcal{I}^c}\right\|_1 - \left\|\boldsymbol{v}^* \boldsymbol{M}_{\mathcal{I}}\right\|_1 = \sum_{i \in \mathcal{I}^c} \left|\boldsymbol{v}^* \boldsymbol{m}_i\right| - \sum_{i \in \mathcal{I}} \left|\boldsymbol{v}^* \boldsymbol{m}_i\right|,$$

namely as a sum of independent random variables. Since $|\mathcal{I}| \le 9n_2\theta/8$, we have

$$\mathbb{E}\left[T\left(\boldsymbol{v}\right)\right] \ge \left(n_2 - \frac{9}{8}\theta n_2 - \frac{9}{8}\theta n_2\right)\mathbb{E}\left[\left|\boldsymbol{v}^*\boldsymbol{m}_1\right|\right] = \left(1 - \frac{9}{4}\theta\right)n_2\mathbb{E}\left[\left|\boldsymbol{v}^*\boldsymbol{m}_1\right|\right] \ge \frac{1}{4}n_2\mathbb{E}\left[\left|\boldsymbol{v}^*\boldsymbol{m}_1\right|\right],$$

where the expectation $\mathbb{E}\left[\left|\boldsymbol{v}^*\boldsymbol{m}_1\right|\right]$ can be lower bounded as

$$\begin{aligned}\mathbb{E}\left[\left|\boldsymbol{v}^*\boldsymbol{m}_1\right|\right] &= \sum_{k=0}^{n_1}\theta^k\left(1-\theta\right)^{n_1-k}\sum_{\mathcal{J}\in\binom{[n_1]}{k}}\mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left|\boldsymbol{v}_{\mathcal{J}}^*\boldsymbol{g}\right|\right]\\ &= \sum_{k=0}^{n_1}\theta^k\left(1-\theta\right)^{n_1-k}\sum_{\mathcal{J}\in\binom{[n_1]}{k}}\sqrt{\frac{2}{\pi}}\left\|\boldsymbol{v}_{\mathcal{J}}\right\| \ge \sqrt{\frac{2}{\pi}}\left\|\mathbb{E}_{\mathcal{J}}\left[\boldsymbol{v}_{\mathcal{J}}\right]\right\| = \sqrt{\frac{2}{\pi}}\theta.\end{aligned}$$

Moreover, by Lemma 9.4 and Lemma B.6, for any $i \in [n_2]$ and any integer $m \ge 2$,

$$\mathbb{E}\left[\left|\boldsymbol{v}^*\boldsymbol{m}_i\right|^m\right] \le \mathbb{E}_{\boldsymbol{Z}\sim\mathcal{N}(0,1)}\left[\left|Z\right|^m\right] \le (m-1)!! \le \frac{m!}{2}.$$

So invoking the moment-control Bernstein's inequality in Lemma A.1, we obtain

$$\mathbb{P}\left[T\left(\boldsymbol{v}\right) < \frac{n_2}{4}\sqrt{\frac{2}{\pi}}\theta - t\right] \le \mathbb{P}\left[T\left(\boldsymbol{v}\right) < \mathbb{E}\left[T\left(\boldsymbol{v}\right)\right] - t\right] \le \exp\left(-\frac{t^2}{2n_2+2t}\right).$$

Taking $t = \frac{n_2}{20}\sqrt{\frac{2}{\pi}}\theta$ and simplifying, we obtain that

$$\mathbb{P}\left[T\left(\boldsymbol{v}\right) < \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta\right] \le \exp\left(-c_1\theta^2 n_2\right) \tag{11.0.1}$$

for some positive constant c_1. Fix $\varepsilon = \sqrt{\frac{2}{\pi}}\frac{\theta}{120}\left[n_1\log\left(n_1 n_2\right)\right]^{-1/2} < 1$. The unit sphere $\mathbb{S}^{n_1-1}$ has an ε-net N_ε of cardinality at most $(3/\varepsilon)^{n_1}$. Consider the event

$$\mathcal{E}_{bg} \doteq \left\{T\left(\boldsymbol{v}\right) \ge \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta\ \ \forall\, \boldsymbol{v}\in N_\varepsilon\right\}.$$

A simple union bound implies

$$\mathbb{P}\left[\mathcal{E}_{bg}^c\right] \le \exp\left(-c_1\theta^2 n_2 + n_1\log\left(\frac{3}{\varepsilon}\right)\right) \le \exp\left(-c_1\theta^2 n_2 + c_2 n_1\log\frac{n_1\log n_2}{\theta}\right), \tag{11.0.2}$$

where $c_2 > 0$ is numerical. Conditioned on $\mathcal{E}_{bg}$, we have that any $\boldsymbol{z}\in\mathbb{S}^{n_1-1}$ can be written as $\boldsymbol{z} = \boldsymbol{v} + \boldsymbol{e}$ for some $\boldsymbol{v}\in N_\varepsilon$ and $\|\boldsymbol{e}\| \le \varepsilon$. Moreover,

$$\begin{aligned}T\left(\boldsymbol{z}\right) &= \left\|\left(\boldsymbol{v}+\boldsymbol{e}\right)^*\boldsymbol{M}_{\mathcal{I}^c}\right\|_1 - \left\|\left(\boldsymbol{v}+\boldsymbol{e}\right)^*\boldsymbol{M}_{\mathcal{I}}\right\|_1 \ge T\left(\boldsymbol{v}\right) - \left\|\boldsymbol{e}^*\boldsymbol{M}_{\mathcal{I}^c}\right\|_1 - \left\|\boldsymbol{e}^*\boldsymbol{M}_{\mathcal{I}}\right\|_1\\ &= \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta - \left\|\boldsymbol{e}^*\boldsymbol{M}\right\|_1 = \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta - \sum_{k=1}^{n_2}\left|\boldsymbol{e}^*\boldsymbol{m}_k\right|\end{aligned}$$

$$\geq \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta - \varepsilon \sum_{k=1}^{n_2} \|\boldsymbol{m}_k\| .$$

By Lemma 9.11, with probability at least $1-\theta\left(n_1 n_2\right)^{-7}-\exp\left(-0.3\theta n_1 n_2\right)$, $\|\boldsymbol{M}\|_\infty \leq 4\sqrt{\log\left(n_1 n_2\right)}$. Thus,

$$T\left(\boldsymbol{z}\right) \geq \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta - \sqrt{\frac{2}{\pi}}\frac{\theta}{120}\frac{n_2\sqrt{n_1}4\sqrt{\log\left(n_1 n_2\right)}}{\sqrt{n_1}\sqrt{\log\left(n_1 n_2\right)}} = \frac{n_2}{6}\sqrt{\frac{2}{\pi}}\theta. \tag{11.0.3}$$

Thus, by (11.0.2), it is enough to take $n_2 > C n_1 \log\left(n_1/\theta\right)/\theta^2$ for sufficiently large $C>0$ to make the overall failure probability small enough so that the lower bound (11.0.3) holds. ■

11.1 Proof of Lemma 6.2

Proof The proof is similar to that of [QSW14]. First, let us assume the dictionary $\boldsymbol{A}_0 = \boldsymbol{I}$. W.l.o.g., suppose that the Riemannian TRM algorithm returns a solution $\widehat{\boldsymbol{q}}$, to which $\boldsymbol{e}_n$ is the nearest signed basis vector. Thus, the rounding LP (6.0.1) takes the form:

$$\operatorname{minimize}_{\boldsymbol{q}} \ \|\boldsymbol{q}^*\boldsymbol{X}_0\|_1, \quad \text{subject to} \quad \langle \boldsymbol{r}, \boldsymbol{q}\rangle = 1. \tag{11.1.1}$$

where the vector $\boldsymbol{r} = \widehat{\boldsymbol{q}}$. Next, We will show whenever $\widehat{\boldsymbol{q}}$ is close enough to $\boldsymbol{e}_n$, w.h.p., the above linear program returns $\boldsymbol{e}_n$. Let $\boldsymbol{X}_0 = \left[\overline{\boldsymbol{X}}; \boldsymbol{x}_n^*\right]$, where $\overline{\boldsymbol{X}} \in \mathbb{R}^{(n-1)\times p}$ and $\boldsymbol{x}_n^*$ is the last row of $\boldsymbol{X}_0$. Set $\boldsymbol{q} = [\overline{\boldsymbol{q}}, q_n]$, where $\overline{\boldsymbol{q}}$ denotes the first $n-1$ coordinates of $\boldsymbol{q}$ and q_n is the last coordinate; similarly for $\boldsymbol{r}$. Let us consider a relaxation of the problem (11.1.1),

$$\operatorname{minimize}_{\boldsymbol{q}} \|\boldsymbol{q}^*\boldsymbol{X}_0\|_1, \quad \text{subject to} \quad q_n r_n + \langle \overline{\boldsymbol{q}}, \overline{\boldsymbol{r}}\rangle \geq 1, \tag{11.1.2}$$

It is obvious that the feasible set of (11.1.2) contains that of (11.1.1). So if $\boldsymbol{e}_n$ is the unique optimal solution (UOS) of (11.1.2), it is the UOS of (11.1.1). Suppose $\mathcal{I} = \operatorname{supp}(\boldsymbol{x}_n)$ and define an event $\mathcal{E}_0 = \left\{|\mathcal{I}| \leq \frac{9}{8}\theta p\right\}$. By Hoeffding's inequality, we know that $\mathbb{P}\left[\mathcal{E}_0^c\right] \leq \exp\left(-\theta^2 p/2\right)$. Now conditioned on $\mathcal{E}_0$ and consider a fixed support $\mathcal{I}$. (11.1.2) can be further relaxed as

$$\operatorname{minimize}_{\boldsymbol{q}} \|\boldsymbol{x}_n\|_1 |q_n| - \left\|\overline{\boldsymbol{q}}^*\overline{\boldsymbol{X}}_{\mathcal{I}}\right\|_1 + \left\|\overline{\boldsymbol{q}}^*\overline{\boldsymbol{X}}_{\mathcal{I}^c}\right\|_1, \quad \text{subject to} \quad q_n r_n + \|\overline{\boldsymbol{q}}\| \|\overline{\boldsymbol{r}}\| \geq 1. \tag{11.1.3}$$

The objective value of (11.1.3) lower bounds that of (11.1.2), and are equal when $\boldsymbol{q} = \boldsymbol{e}_n$. So if $\boldsymbol{q} = \boldsymbol{e}_n$ is UOS of (11.1.3), it is UOS of (11.1.1). By Lemma 11.1, we know that

$$\left\|\overline{\boldsymbol{q}}^*\overline{\boldsymbol{X}}_{\mathcal{I}^c}\right\|_1 - \left\|\overline{\boldsymbol{q}}^*\overline{\boldsymbol{X}}_{\mathcal{I}}\right\|_1 \ \geq \ \frac{p}{6}\sqrt{\frac{2}{\pi}}\theta\,\|\overline{\boldsymbol{q}}\|$$

holds w.h.p. when $p \geq C_1(n-1)\log\left((n-1)/\theta\right)/\theta^2$. Let $\zeta = \frac{p}{6}\sqrt{\frac{2}{\pi}}\theta$, thus we can further lower bound the objective value in (11.1.3) by

$$\operatorname{minimize}_{\boldsymbol{q}} \ \|\boldsymbol{x}_n\|_1 |q_n| + \zeta \|\overline{\boldsymbol{q}}\|, \quad \text{subject to} \quad q_n r_n + \|\overline{\boldsymbol{q}}\| \|\overline{\boldsymbol{r}}\| \geq 1. \tag{11.1.4}$$

By similar arguments, if $\boldsymbol{e}_n$ is the UOS of (11.1.4), it is also the UOS of (11.1.1). For the optimal solution of (11.1.4), notice that it is necessary to have $\operatorname{sign}(q_n) = \operatorname{sign}(r_n)$ and $q_n r_n + \|\overline{\boldsymbol{q}}\| \|\overline{\boldsymbol{r}}\| = 1$. Therefore, the problem (11.1.4) is equivalent to

$$\operatorname{minimize}_{q_n} \ \|\boldsymbol{x}_n\|_1 |q_n| + \zeta \frac{1 - |r_n| |q_n|}{\|\overline{\boldsymbol{r}}\|}, \quad \text{subject to} \quad |q_n| \leq \frac{1}{|r_n|}. \tag{11.1.5}$$

Notice that the problem (11.1.5) is a linear program in $|q_n|$ with a compact feasible set, which indicates that the optimal solution only occurs at the boundary points $|q_n| = 0$ and $|q_n| = 1/|r_n|$. Therefore, $\boldsymbol{q} = \boldsymbol{e}_n$ is the UOS of (11.1.5) if and only if

$$\frac{1}{|r_n|} \|\boldsymbol{x}_n\|_1 < \frac{\zeta}{\|\overline{\boldsymbol{r}}\|}. \tag{11.1.6}$$

Conditioned on $\mathcal{E}_0$, by using the Gaussian concentration bound, we have

$$\mathbb{P}\left[\|\boldsymbol{x}_n\|_1 \geq \frac{9}{8}\sqrt{\frac{2}{\pi}}\theta p + t\right] \ \leq \ \mathbb{P}\left[\|\boldsymbol{x}_n\|_1 \geq \mathbb{E}\left[\|\boldsymbol{x}_n\|_1\right] + t\right] \ \leq \ \exp\left(-\frac{t^2}{2p}\right),$$

which means that

$$\mathbb{P}\left[\|\boldsymbol{x}_n\|_1 \geq \frac{5}{4}\sqrt{\frac{2}{\pi}}\theta p\right] \ \leq \ \exp\left(-\frac{\theta^2 p}{64\pi}\right). \tag{11.1.7}$$

Therefore, by (11.1.6) and (11.1.7), for $\boldsymbol{q} = \boldsymbol{e}_n$ to be the UOS of (11.1.1) w.h.p., it is sufficient to have

$$\frac{5}{4|r_n|}\sqrt{\frac{2}{\pi}}\theta p \ < \ \frac{\theta p}{6\sqrt{1 - |r_n|^2}}\sqrt{\frac{2}{\pi}}, \tag{11.1.8}$$

which is implied by

$$|r_n| \ > \ \frac{249}{250}.$$

The failure probability can be estimated via a simple union bound. Since the above argument holds uniformly for any fixed support set $\mathcal{I}$, we obtain the desired result.

When our dictionary $\boldsymbol{A}_0$ is an arbitrary orthogonal matrix, it only rotates the row subspace of $\boldsymbol{X}_0$. Thus, w.l.o.g., suppose the TRM algorithm returns a solution $\widehat{\boldsymbol{q}}$, to which $\boldsymbol{A}_0\boldsymbol{q}_\star$ is the nearest "target" with $\boldsymbol{q}_\star$ a signed basis vector. By a change of variable $\tilde{\boldsymbol{q}} = \boldsymbol{A}_0^* \boldsymbol{q}$, the problem (11.1.1) is of the form

$$\operatorname{minimize}_{\tilde{\boldsymbol{q}}} \left\| \widetilde{\boldsymbol{q}}^* \boldsymbol{X}_0 \right\|_1, \quad \text{subject to} \quad \langle \boldsymbol{A}_0^* \boldsymbol{r}, \tilde{\boldsymbol{q}} \rangle = 1,$$

obviously our target solution for $\tilde{\boldsymbol{q}}$ is again the standard basis $\boldsymbol{q}_\star$. By a similar argument above, we only need $\langle \boldsymbol{A}_0^* \boldsymbol{r}, \boldsymbol{e}_n \rangle > 249/250$ to exactly recover the target, which is equivalent to $\langle \boldsymbol{r}, \widehat{\boldsymbol{q}}_\star \rangle > 249/250$. This implies that our rounding (6.0.1) is invariant to change of basis, completing the proof. ■

11.2 Proof of Lemma 6.4

Proof Define $\widetilde{\boldsymbol{q}} \doteq (\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi})^* \boldsymbol{q}$. By Lemma 4.14, and in particular (4.3.2), when

$$p \geq \frac{C}{c_\star^2 \theta} \max \left\{ \frac{n^4}{\mu^4}, \frac{n^5}{\mu^2} \right\} \kappa^8 \left(\boldsymbol{A}_0 \right) \log^4 \left(\frac{\kappa \left(\boldsymbol{A}_0 \right) n}{\mu \theta} \right),$$

$\|\boldsymbol{\Xi}\| \leq 1/2$ so that $\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi}$ is invertible. Then the LP rounding can be written as

$$\operatorname{minimize}_{\widetilde{\boldsymbol{q}}} \left\| \widetilde{\boldsymbol{q}}^* \boldsymbol{X}_0 \right\|_1, \quad \text{subject to} \quad \left\langle (\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi})^{-1} \boldsymbol{r}, \widetilde{\boldsymbol{q}} \right\rangle = 1.$$

By Lemma 6.2, to obtain $\widetilde{\boldsymbol{q}} = \boldsymbol{e}_n$ from this LP, it is enough to have

$$\left\langle (\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi})^{-1} \boldsymbol{r}, \boldsymbol{e}_n \right\rangle \geq 249/250,$$

and $p \geq C n^2 \log(n/\theta)/\theta$ for some large enough C. This implies that to obtain $\boldsymbol{q}_\star$ for the original LP, such that $(\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi})^* \boldsymbol{q}_\star = \boldsymbol{e}_n$, it is enough that

$$\left\langle (\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi})^{-1} \boldsymbol{r}, (\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi})^* \boldsymbol{q}_\star \right\rangle = \langle \boldsymbol{r}, \boldsymbol{q}_\star \rangle \geq 249/250,$$

completing the proof. ■

11.3 Proof of Lemma 6.5

Proof Note that $[\boldsymbol{q}_\star^1, \dots, \boldsymbol{q}_\star^\ell] = (\boldsymbol{Q}^* + \boldsymbol{\Xi}^*)^{-1} [\boldsymbol{e}_1, \dots, \boldsymbol{e}_\ell]$, we have

$$\begin{aligned} \boldsymbol{U}^* (\boldsymbol{Q} + \boldsymbol{\Xi}) \boldsymbol{X}_0 &= \boldsymbol{U}^* (\boldsymbol{Q}^* + \boldsymbol{\Xi}^*)^{-1} (\boldsymbol{Q} + \boldsymbol{\Xi})^* (\boldsymbol{Q} + \boldsymbol{\Xi}) \boldsymbol{X}_0 \\ &= \boldsymbol{U}^* \left[\boldsymbol{q}_\star^1, \dots, \boldsymbol{q}_\star^\ell \mid \widehat{\boldsymbol{V}} \right] (\boldsymbol{I} + \boldsymbol{\Delta}_1) \boldsymbol{X}_0, \end{aligned}$$

where $\widehat{\boldsymbol{V}} \doteq (\boldsymbol{Q}^* + \boldsymbol{\Xi}^*)^{-1}[\boldsymbol{e}_{\ell+1}, \dots, \boldsymbol{e}_n]$, and the matrix $\boldsymbol{\Delta}_1 = \boldsymbol{Q}^*\boldsymbol{\Xi} + \boldsymbol{\Xi}^*\boldsymbol{Q} + \boldsymbol{\Xi}^*\boldsymbol{\Xi}$ so that $\|\boldsymbol{\Delta}_1\| \le 3\|\boldsymbol{\Xi}\|$. Since $\boldsymbol{U}^*\left[\boldsymbol{q}_\star^1, \dots, \boldsymbol{q}_\star^\ell \mid \widehat{\boldsymbol{V}}\right] = \left[\boldsymbol{0} \mid \boldsymbol{U}^*\widehat{\boldsymbol{V}}\right]$, we have

$$\boldsymbol{U}^*(\boldsymbol{Q} + \boldsymbol{\Xi})\boldsymbol{X}_0 = \left[\boldsymbol{0} \mid \boldsymbol{U}^*\widehat{\boldsymbol{V}}\right]\boldsymbol{X}_0 + \left[\boldsymbol{0} \mid \boldsymbol{U}^*\widehat{\boldsymbol{V}}\right]\boldsymbol{\Delta}_1\boldsymbol{X}_0 = \boldsymbol{U}^*\widehat{\boldsymbol{V}}\boldsymbol{X}_0^{[n-\ell]} + \boldsymbol{\Delta}_2\boldsymbol{X}_0, \tag{11.3.1}$$

where $\boldsymbol{\Delta}_2 = \left[\boldsymbol{0} \mid \boldsymbol{U}^*\widehat{\boldsymbol{V}}\right]\boldsymbol{\Delta}_1$. Let $\delta = \|\boldsymbol{\Xi}\|$, so that

$$\|\boldsymbol{\Delta}_2\| \le \frac{\|\boldsymbol{\Delta}_1\|}{\sigma_{\min}(\boldsymbol{Q} + \boldsymbol{\Xi})} \le \frac{3\|\boldsymbol{\Xi}\|}{\sigma_{\min}(\boldsymbol{Q} + \boldsymbol{\Xi})} \le \frac{3\delta}{1-\delta}. \tag{11.3.2}$$

Since the matrix $\widehat{\boldsymbol{V}}$ is near orthogonal, it can be decomposed as $\widehat{\boldsymbol{V}} = \boldsymbol{V} + \boldsymbol{\Delta}_3$, where $\boldsymbol{V}$ is orthogonal, and $\boldsymbol{\Delta}_3$ is a small perturbation. Obviously, $\boldsymbol{V} = \boldsymbol{U}\boldsymbol{R}$ for some orthogonal matrix $\boldsymbol{R}$, so that spans the same subspace as that of $\boldsymbol{U}$. Next, we control the spectral norm of $\boldsymbol{\Delta}_3$ so that it is sufficiently small,

$$\|\boldsymbol{\Delta}_3\| = \min_{\boldsymbol{R} \in O_\ell}\left\|\boldsymbol{U}\boldsymbol{R} - \widehat{\boldsymbol{V}}\right\| \le \min_{\boldsymbol{R} \in O_\ell}\left\|\boldsymbol{U}\boldsymbol{R} - \boldsymbol{Q}_{[n-\ell]}\right\| + \left\|\boldsymbol{Q}_{[n-\ell]} - \widehat{\boldsymbol{V}}\right\|, \tag{11.3.3}$$

where $\boldsymbol{Q}_{[n-\ell]}$ collects the last $n-\ell$ columns of $\boldsymbol{Q}$, i.e., $\boldsymbol{Q} = [\boldsymbol{Q}_{[\ell]}, \boldsymbol{Q}_{[n-\ell]}]$. To bound the second term on the right, we have

$$\left\|\boldsymbol{Q}_{[n-\ell]} - \widehat{\boldsymbol{V}}\right\| \le \left\|\boldsymbol{Q}^{-1} - (\boldsymbol{Q} + \boldsymbol{\Xi})^{-1}\right\| \le \frac{\|\boldsymbol{Q}^{-1}\|\,\|\boldsymbol{Q}^{-1}\boldsymbol{\Xi}\|}{1 - \|\boldsymbol{Q}^{-1}\boldsymbol{\Xi}\|} \le \frac{\delta}{1-\delta},$$

where we have used perturbation bound for matrix inverse (see, e.g., Theorem 2.5 of Chapter III in [SS90]). To bound the first term, from Lemma B.13, it is enough to upper bound the largest principal angle θ_1 between the subspaces $\mathrm{span}([\boldsymbol{q}_\star^1, \dots, \boldsymbol{q}_\star^\ell])$, and that spanned by $\boldsymbol{Q}[\boldsymbol{e}_1, \dots, \boldsymbol{e}_\ell]$. Write $\boldsymbol{I}_{[\ell]} \doteq [\boldsymbol{e}_1, \dots, \boldsymbol{e}_\ell]$ for short, we bound $\sin\theta_1$ as

$$\begin{aligned}
\sin\theta_1 &\le \left\|\boldsymbol{Q}\boldsymbol{I}_{[\ell]}\boldsymbol{I}_{[\ell]}^*\boldsymbol{Q}^* - (\boldsymbol{Q}^* + \boldsymbol{\Xi}^*)^{-1}\boldsymbol{I}_{[\ell]}\left(\boldsymbol{I}_{[\ell]}^*(\boldsymbol{Q} + \boldsymbol{\Xi})^{-1}(\boldsymbol{Q}^* + \boldsymbol{\Xi}^*)^{-1}\boldsymbol{I}_{[\ell]}\right)^{-1}\boldsymbol{I}_{[\ell]}^*(\boldsymbol{Q} + \boldsymbol{\Xi})^{-1}\right\| \\
&= \left\|\boldsymbol{Q}\boldsymbol{I}_{[\ell]}\boldsymbol{I}_{[\ell]}^*\boldsymbol{Q}^* - (\boldsymbol{Q}^* + \boldsymbol{\Xi}^*)^{-1}\boldsymbol{I}_{[\ell]}\left(\boldsymbol{I}_{[\ell]}^*(\boldsymbol{I} + \boldsymbol{\Delta}_1)^{-1}\boldsymbol{I}_{[\ell]}\right)^{-1}\boldsymbol{I}_{[\ell]}^*(\boldsymbol{Q} + \boldsymbol{\Xi})^{-1}\right\| \\
&\le \left\|\boldsymbol{Q}\boldsymbol{I}_{[\ell]}\boldsymbol{I}_{[\ell]}^*\boldsymbol{Q}^* - (\boldsymbol{Q}^* + \boldsymbol{\Xi}^*)^{-1}\boldsymbol{I}_{[\ell]}\boldsymbol{I}_{[\ell]}^*(\boldsymbol{Q} + \boldsymbol{\Xi})^{-1}\right\| \\
&\qquad + \left\|(\boldsymbol{Q}^* + \boldsymbol{\Xi}^*)^{-1}\boldsymbol{I}_{[\ell]}\left[\boldsymbol{I} - \left(\boldsymbol{I}_{[\ell]}^*(\boldsymbol{I} + \boldsymbol{\Delta}_1)^{-1}\boldsymbol{I}_{[\ell]}\right)^{-1}\right]\boldsymbol{I}_{[\ell]}^*(\boldsymbol{Q} + \boldsymbol{\Xi})^{-1}\right\| \\
&\le \left(1 + \frac{1}{\sigma_{\min}(\boldsymbol{Q} + \boldsymbol{\Xi})}\right)\left\|\boldsymbol{Q}^{-1} - (\boldsymbol{Q} + \boldsymbol{\Xi})^{-1}\right\| + \frac{1}{\sigma_{\min}^2(\boldsymbol{Q} + \boldsymbol{\Xi})}\left\|\boldsymbol{I} - \left(\boldsymbol{I}_{[\ell]}^*(\boldsymbol{I} + \boldsymbol{\Delta}_1)^{-1}\boldsymbol{I}_{[\ell]}\right)^{-1}\right\| \\
&\le \left(1 + \frac{1}{1-\delta}\right)\frac{\delta}{1-\delta} + \frac{1}{(1-\delta)^2}\frac{\left\|\boldsymbol{I}_{[\ell]}^*(\boldsymbol{I} + \boldsymbol{\Delta}_1)^{-1}\boldsymbol{I}_{[\ell]} - \boldsymbol{I}\right\|}{1 - \left\|\boldsymbol{I}_{[\ell]}^*(\boldsymbol{I} + \boldsymbol{\Delta}_1)^{-1}\boldsymbol{I}_{[\ell]} - \boldsymbol{I}\right\|} \\
&\le \left(1 + \frac{1}{1-\delta}\right)\frac{\delta}{1-\delta} + \frac{1}{(1-\delta)^2}\frac{\|\boldsymbol{\Delta}_1\|}{1 - 2\|\boldsymbol{\Delta}_1\|},
\end{aligned}$$

where in the first line we have used the fact that for any full column rank matrix $\boldsymbol{M}$, $\boldsymbol{M}(\boldsymbol{M}^*\boldsymbol{M})^{-1}\boldsymbol{M}^*$ is the orthogonal projection onto the its column span, and to obtain the fifth and six lines we have invoked the matrix inverse perturbation bound again. Use the facts that $\delta < 1/20$ and $\|\boldsymbol{\Delta}_1\| \le 3\delta < 1/2$, we have

$$\sin\theta_1 \le \frac{(2-\delta)\delta}{(1-\delta)^2} + \frac{3\delta}{(1-\delta)^2(1-6\delta)} = \frac{5\delta - 13\delta^2 + 6\delta^3}{(1-\delta)^2(1-6\delta)} \le 8\delta.$$

For $\delta < 1/20$, the upper bound is nontrivial. By Lemma B.13,

$$\min_{\boldsymbol{R}\in O_\ell} \left\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{Q}_{[n-\ell]} \right\| \le \sqrt{2 - 2\cos\theta_1} \le \sqrt{2 - 2\cos^2\theta_1} = \sqrt{2}\sin\theta_1 \le 8\sqrt{2}\delta.$$

Put the estimates above, there exists an orthogonal matrix $\boldsymbol{R} \in O_\ell$ such that $\boldsymbol{V} = \boldsymbol{U}\boldsymbol{R}$ and $\widehat{\boldsymbol{V}} = \boldsymbol{V} + \boldsymbol{\Delta}_3$ with

$$\|\boldsymbol{\Delta}_3\| \le \delta/(1-\delta) + 8\sqrt{2}\delta \le 12.5\delta. \tag{11.3.4}$$

Therefore, by (11.3.1), we obtain

$$\boldsymbol{U}^*(\boldsymbol{Q} + \boldsymbol{\Xi})\boldsymbol{X}_0 = \boldsymbol{U}^*\boldsymbol{V}\boldsymbol{X}_0^{[n-\ell]} + \boldsymbol{\Delta}, \quad \text{with} \quad \boldsymbol{\Delta} \doteq \boldsymbol{U}^*\boldsymbol{\Delta}_3\boldsymbol{X}_0^{[n-\ell]} + \boldsymbol{\Delta}_2\boldsymbol{X}_0. \tag{11.3.5}$$

By using the results in (11.3.2) and (11.3.4), we get the desired result. ■

Part III

Generalized Phase Retrieval

Can we recover a complex signal from its Fourier magnitudes? More generally, given a set of m nonlinear measurements $y_k = |\boldsymbol{a}_k^* \boldsymbol{x}|$ for $k = 1, \dots, m$, is it possible to recover $\boldsymbol{x} \in \mathbb{C}^n$ (i.e., length-n complex vector)? This *generalized phase retrieval* (GPR) problem is a fundamental task in various disciplines, and has been the subject of much recent investigation. Natural nonconvex methods often work remarkably well for GPR in practice, but lack clear theoretical explanations. In this part, we take a step towards bridging this gap. We prove that when the measurement vectors $\boldsymbol{a}_k$'s are generic (i.i.d. complex Gaussian) and the number of measurements is large enough ($m \geq Cn \log^3 n$), with high probability, a natural least-squares formulation for GPR has the following benign geometric structure: (1) there are no spurious local minimizers, and all global minimizers are equal to the target signal $\boldsymbol{x}$, up to a global phase; and (2) the objective function has a negative curvature around each saddle point. In other words, the least-squares formulation under study lies in the $\mathcal{X}$ family. This structure allows a number of iterative optimization methods to efficiently find a global minimizer, without special initialization. To corroborate the claim, we describe and analyze a second-order trust-region algorithm.

This part is organized as follows. We provide background on the GPR problem and an overview of the geometric structure of the least-squares formulation in Chapter 12. In Chapter 13, we provide a quantitative characterization of the global geometry for GPR and highlight main technical challenges in establishing the results. Based on this characterization, in Chapter 14 we present a modified trust-region method for minimizing the least-squares formulation from an arbitrary initialization, which leads to our main computational guarantee. In Chapter 15 we study the empirical performance of our method for GPR. Chapter 16 concludes the main body with a discussion of open problems. Chapter 17 and Chapter 18 collect detailed proofs to technical results for the geometric analysis and algorithmic analysis, respectively.

This part is based on our technical report:

A Geometric Analysis of Phase Retrieval. http://arxiv.org/abs/1602.06664

The codes to reproduce all the figures and the experimental results can be found online:

https://github.com/sunju/pr_plain

Chapter 12

Introduction

Practical application is found by not looking for it, and one can say that the whole progress of civilization rests on that principle.

Jacques Hadamard

12.1 Generalized phase retrieval and a nonconvex formulation

This part concerns the problem of recovering an n-dimensional complex vector $\boldsymbol{x}$ from the magnitude $y_k = |\boldsymbol{a}_k^* \boldsymbol{x}|$ of its projections onto a collection of known complex vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_m \in \mathbb{C}^n$. Obviously, one can only hope to recover $\boldsymbol{x}$ up to a global phase, as $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}$ for all $\phi \in [0, 2\pi)$ gives exactly the same set of measurements. The *generalized phase retrieval* problem asks whether it is possible to recover $\boldsymbol{x}$, up to this fundamental ambiguity:

> **Generalized Phase Retrieval Problem**: Is it possible to *efficiently* recover an unknown $\boldsymbol{x}$ from $y_k = |\boldsymbol{a}_k^* \boldsymbol{x}|$ $(k = 1, \dots, m)$, up to a global phase factor $\mathrm{e}^{\mathrm{i}\phi}$?

This problem has attracted substantial recent interest, due to its connections to fields such as crystallography, optical imaging and astronomy. In these areas, one often has access only to the Fourier magnitudes of a complex signal $\boldsymbol{x}$, i.e., $|\mathcal{F}(\boldsymbol{x})|$ [Mil90, Rob93, Wal63, DF87]. The phase information is hard or infeasible to record due to physical constraints. The problem of recovering the signal $\boldsymbol{x}$ from its Fourier magnitudes $|\mathcal{F}(\boldsymbol{x})|$ is naturally termed (Fourier) phase retrieval (PR). It is easy to see PR as a special version of GPR, with the $\boldsymbol{a}_k$'s the Fourier basis vectors. GPR also sees applications in electron microscopy [MIJ$^+$02], diffraction and array imaging [BDP$^+$07, CMP11], acoustics [BCE06, Bal10], quantum mechanics [Cor06, Rei65] and quantum information [HMW13]. We refer the reader to survey papers [SEC+15, JEH15] for accounts of recent developments in the theory, algorithms, and applications of GPR.

For GPR, simple or even heuristic methods based on nonconvex optimization often work surprisingly well in practice (e.g., [Fie82, GS72], and many more cited in [SEC+15, JEH15]). However, investigation into provable recovery methods, particularly based on nonconvex optimization, has started only relatively recently [NJS13, CESV13, CSV13, CL14, CLS15a, WdM15, VX14, ABFM14, CLS15b, CC15, WWS15]. The surprising effectiveness of simple methods on GPR remains largely mysterious. In this part, we take a step towards bridging this gap.

We focus on a natural least-squares formulation[1] – discussed systematically in [SEC+15, JEH15] and studied theoretically in [CLS15b, WWS15],

$$\mathrm{minimize}_{\boldsymbol{z}\in\mathbb{C}^n}\ f(\boldsymbol{z}) \doteq \frac{1}{2m}\sum_{k=1}^{m}\left(y_k^2 - |\boldsymbol{a}_k^*\boldsymbol{z}|^2\right)^2. \tag{12.1.1}$$

We assume the $\boldsymbol{a}_k$'s are independent identically distributed (i.i.d.) complex Gaussian:

$$\boldsymbol{a}_k = \frac{1}{\sqrt{2}}\left(X_k + \mathrm{i}Y_k\right), \text{ with } X_k, Y_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n) \text{ independent.} \tag{12.1.2}$$

$f(\boldsymbol{z})$ is a fourth-order polynomial in $\boldsymbol{z}$, and is nonconvex. A-priori, there is little reason to believe that simple iterative methods can solve this problem without special initialization. Typical local convergence (i.e., convergence to a local minimizer) guarantees in optimization require an initialization near the target minimizer [Ber99]. Moreover, existing results on provable recovery using (12.1.1) and related formulations rely on careful initialization in the vicinity of the ground truth [NJS13, CLS15b, CC15, WWS15].

12.2 A curious experiment

We apply gradient descent to $f(\boldsymbol{z})$, starting from a *random initialization* $\boldsymbol{z}^{(0)}$:

$$\boldsymbol{z}^{(r+1)} = \boldsymbol{z}^{(r)} - \mu\nabla_{\boldsymbol{z}} f(\boldsymbol{z}^{(r)}),$$

[1] Another least-squares formulation, $\mathrm{minimize}_{\boldsymbol{z}}\ \frac{1}{2m}\sum_{k=1}^{m}(y_k - |\boldsymbol{a}_k^*\boldsymbol{z}|)^2$, was studied in the seminal works [Fie82, GS72]. An obvious advantage of the $f(\boldsymbol{z})$ studied here is that it is differentiable – not in the usual complex calculus sense, but in the Wirtinger calculus sense; see Section 12.5 for a brief review of Wirtinger calculus.

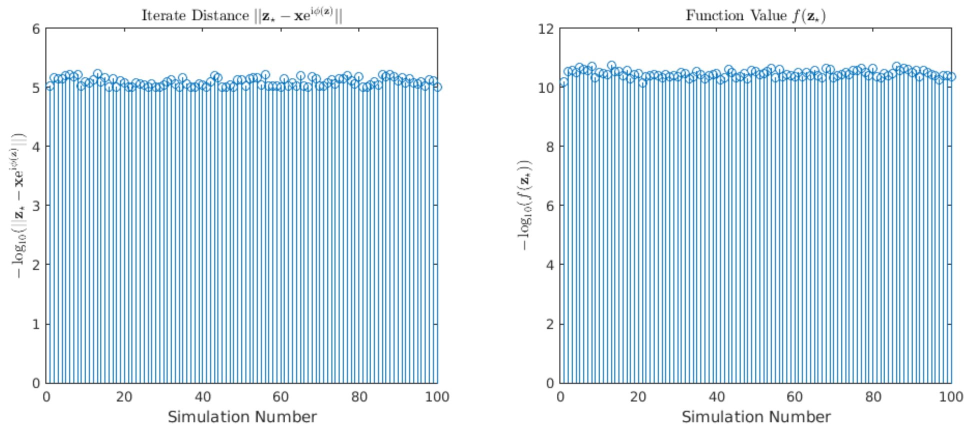


Figure 12.1: Gradient descent with random initialization seems to always return a global solution for (12.1.1)! Here $n = 100$, $m = 5n \log n$, step size $\mu = 0.05$, and stopping criterion is $\|\nabla_{\boldsymbol{z}} f(\boldsymbol{z})\| \le 10^{-5}$. We fix the set of random measurements and the ground-truth signal $\boldsymbol{x}$. The experiments are repeated for 100 times with independent random initializations. $\boldsymbol{z}_\star$ denotes the final iterate at convergence. (Left) Final distance to the target; (Right) Final function value (0 if globally optimized). Both vertical axes are on $-\log_{10}(\cdot)$ scale.

where the step size μ is fixed for simplicity[2]. The result is quite striking (Figure 12.1): for a fixed problem instance (fixed set of random measurements and fixed target $\boldsymbol{x}$), gradient descent seems to always return a *global minimizer* (i.e., the target $\boldsymbol{x}$ up to a global phase shift), across many independent random initializations! This contrasts with the typical "mental picture" of nonconvex objectives as possessing many spurious local minimizers.

12.3 A geometric analysis

The numerical surprise described above is not completely isolated. Simple methods have been observed to work surprisingly well for practical PR [Fie82, GS72, SEC$^+$15, JEH15]. In this work, we take a step towards explaining this phenomenon. We show that *although the function* (12.1.1) *is nonconvex, when m is reasonably large, it actually has benign global geometry (i.e., $\mathcal{X}$-ness) which allows it to be globally optimized by efficient iterative methods, regardless of the initialization.*

This geometric structure is evident for GPR in $\mathbb{R}^2$. Figure 12.2 plots the function landscape of $f(\boldsymbol{z})$ for this case with large m (i.e., think of $\mathbb{E}_{\boldsymbol{a}}[f(\boldsymbol{z})]$). Notice that (i) the only local minimizers are exactly $\pm\boldsymbol{x}$ – they are also global minimizers;[3] (ii) there are saddle points (and a local maximizer), but around them there is

[2] Here the gradient is defined based on the Wirtinger derivatives [KD09]; see also [CLS15b]. This notion of gradient is a natural choice when optimizing real-valued functions of complex variables. For better speed, our implementation of the gradient descent method is actually a line-search variant.

[3] Note that the global sign cannot be recovered.

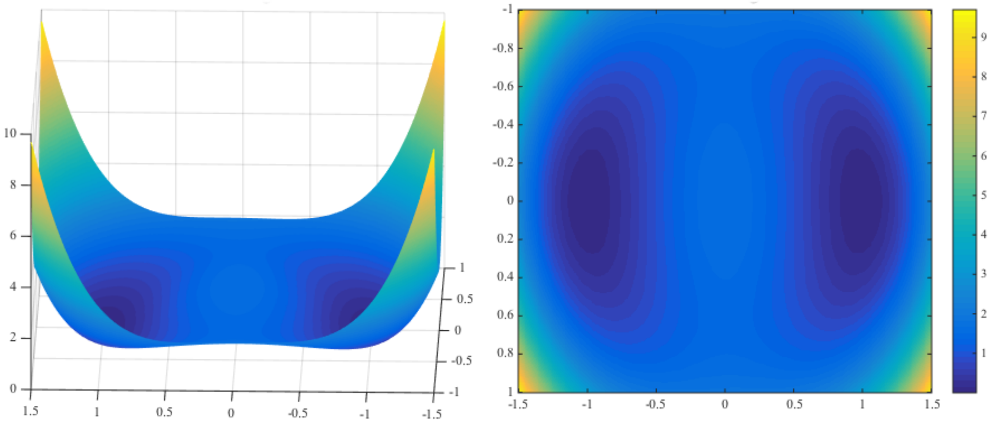


Figure 12.2: Function landscape of (12.1.1) for $\boldsymbol{x} = [1; 0]$ and $m \to \infty$. The only local and also global minimizers are $\pm\boldsymbol{x}$. There are two saddle points near $\pm[0; 1/\sqrt{2}]$, around each there is a negative curvature direction along $\pm\boldsymbol{x}$. (Left) The function graph; (Right) The same function visualized as a color image. The measurement vectors $\boldsymbol{a}_k$'s are taken as i.i.d. standard real Gaussian in this version.

negative curvature in the $\pm\boldsymbol{x}$ direction. Intuitively, any algorithm that can successfully escape from this kind of saddle point (and local maximizer) can in fact find a global minimizer, i.e., recover the target signal $\boldsymbol{x}$.

We prove that an analogous geometric structure exists, with high probability (w.h.p.)[4], for GPR in $\mathbb{C}^n$, when m is reasonably large (Theorem 13.2). In particular, we show that when $m \geq Cn\log^3 n$, w.h.p., (i) the only local and also global minimizers to (12.1.1) are the target $\boldsymbol{x}e^{\mathrm{i}\phi}$ for $\phi \in [0, 2\pi)$; (ii) at any point in $\mathbb{C}^n$, either the gradient is large, or the curvature is negative in a certain direction, or it is near a minimizer. Moreover, in the vicinity of the minimizers, on the orthogonal complement of a single flat direction (which occurs because $f(\boldsymbol{z}e^{\mathrm{i}\phi}) = f(\boldsymbol{z})$ for every $\boldsymbol{z}$, ϕ), the objective function is strongly convex. In other words, under conditions on m, $f(\boldsymbol{z})$ is an $\mathcal{X}$ function modulo the flat direction at each point.

Because of this global geometry, a wide range of efficient iterative methods can obtain a global minimizer to $f(\boldsymbol{z})$, regardless of initialization. Examples include the noisy gradient and stochastic gradient methods [GHJY15] (see also [LSJR16]), curvilinear search [Gol80] and trust-region methods [CGT00, NP06, SQW15b]. The key property that the methods must possess is the ability to escape from saddle points (and local maximizers) at which the Hessian has a strictly negative eigenvalue.

We corroborate this claim by developing a second-order trust-region method for this problem, and prove that (Theorem 14.10) (i) from any initialization, it efficiently obtains a close approximation (i.e., up to numerical precision) of the target $\boldsymbol{x}$ (up to a global phase) and (ii) it exhibits quadratic convergence in the vicinity of the global minimizers.

[4] The probability is with respect to drawing of $\boldsymbol{a}_k$'s.

In sum, our geometric analysis produces the following result.

> **Theorem 12.1 (Informal Statement of Our Main Results; See Theorems 13.2 and 14.10.)** *When $m \geq Cn \log^3 n$, with probability at least $1 - cm^{-1}$, the function $f(\boldsymbol{z})$ has no spurious local minimizers. The only global minimizers are the target $\boldsymbol{x}$ and its equivalent copies, and at all saddle points the function has directional negative curvature. Moreover, with at least the same probability, the trust-region method with properly set step size parameter find a global minimizer of $f(\boldsymbol{z})$ in polynomial time, from an arbitrary initialization in the zero-centered complex ball with radius $R_0 \doteq 3(\frac{1}{m}\sum_{k=1}^{m} y_k^2)^{1/2}$. Here C and c are positive absolute constants.*

The choice of R_0 above allows us to state a result with a concise bound on the number of iterations required to converge. However, under our probability model, w.h.p., the trust-region method succeeds from any initialization. There are two caveats to this claim. First, one must choose the parameters of the method appropriately. Second, the number of iterations depends on how far away from the truth the method starts.

Our results asserts that when the $\boldsymbol{a}_k$'s are *numerous* and *generic* enough, GPR can be solved efficiently. Similar conclusions have been obtained in [NJS13, CLS15b, CC15, WWS15]. One salient feature of our result is that the optimization method is "initialization free" - any initialization in the prescribed ball works, while all prior methods [NJS13, CLS15b, CC15, WWS15] require careful initializations that are already near the unknown target $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}$. This distinctive property follows directly from the benign global geometry of $f(\boldsymbol{z})$. We believe that this sheds light on mechanism of the above numerical surprise.

12.4 Prior arts and connections

The survey papers [SEC$^+$15, JEH15] provide accounts of recent progress on GPR. In this section, we focus on provable efficient (nonconvex) methods for GPR, and draw connections to other work on provable nonconvex heuristics for practical problems.

Provable methods for GPR. Although simple methods for GPR have been used effectively in practice [GS72, Fie82, SEC$^+$15, JEH15], only recently have researchers begun to develop methods with provable performance guarantees. The first results of this nature were obtained using semidefinite programming (SDP) relaxations [CESV13, CSV13, CL14, CLS15a, WdM15, VX14]. While this represented a substantial advance in theory, the computational complexity of semidefinite programming limits the practicality of this approach.[5]

[5] Another line of research [BCE06, BBCE09, ABFM14] seeks to co-design the measurements and recovery algorithms based on frame- or graph-theoretic tools.

Recently, several provable *nonconvex* methods have been proposed for GPR. [NJS13] augmented the seminal error-reduction method [GS72] with spectral initialization and resampling to obtain the first provable nonconvex method for GPR. [CLS15b] studied the nonconvex formulation (12.1.1) under the same hypotheses as this paper, and showed that a combination of spectral initialization and local gradient descent recovers the true signal with near-optimal sample complexity. [CC15] worked with a different nonconvex formulation, and refined the spectral initialization and the local gradient descent with a step-adaptive truncation. With the modifications, they reduced the sample requirement to the optimal order.[6] Compared to the SDP-based methods, these methods are more scalable and closer to methods used in practice. All three analyses are local in nature, and depend on the spectral initializer being sufficiently close to the target signal.

In contrast, we explicitly characterize the global function landscape of (12.1.1). Its benign geometric structure allows several algorithmic choices that need *no special initialization*. In fact, the spectral initialization used in [CLS15b] lands the iterate sequence in the region in which the objective is (restrictedly) strongly convex ($\mathcal{R}_3$ in Theorem 13.2). The analysis of [CLS15b] is based on a property that ensures the gradient descent method is locally contractive near the target set, which is closely linked to (restricted) local convexity. [Sol14] and [WWS15] explicitly established local strong convexity near the target set for GPR in $\mathbb{R}^n$.

Geometric analysis of other nonconvex problems. The approach taken here is similar in spirit to our geometric analysis of the nonconvex formulation for complete dictionary learning in Part II (see also [SQW15a]). Particularly, we show that the nonconvex formulations studied are $\mathcal{X}$ functions in appropriate sense. Despite these similarities, GPR raises several novel technical challenges: the objective is heavy-tailed, and minimizing the number of measurements is important.

Our work sits amid the recent surge of work on provable nonconvex methods for practical problems. Besides GPR studied here, this line of work includes low-rank matrix recovery [KMO10, JNS13, Har14, HW14, NNS$^+$14, JN14, SL14, WCCL15, SRO15, ZL15, TBSR15, CW15], tensor recovery [JO14, AGJ14a, AGJ14b, AJSN15, GHJY15], structured element pursuit [QSW14, HSSS15], dictionary learning [AAJ$^+$13, AGM13, AAN13, ABGM14, AGMM15, SQW15a], mixed regression [YCS13, SA14c], blind deconvolution [LWB13, LJ15, LLJB15], super resolution [EW15], phase synchronization [Bou16], numerical linear algebra [JJKN15], and so forth. Most of the methods adopt the strategy of initialization plus local refinement we alluded to above. In contrast, our geometric analysis allows flexible algorithm design (i.e., separation of geometry and

[6] In addition, [CC15] shows that the measurements can be non-adaptive, in the sense that a single, randomly chosen collection of vectors $\boldsymbol{a}_i$ can simultaneously recover every $\boldsymbol{x} \in \mathbb{C}^n$. Results in [NJS13, CLS15b] and this paper pertain only to adaptive measurements that recover any fixed signal $\boldsymbol{x}$ with high probability.

algorithms) and gives some clues to the underlying mechanism of the behaviors of nonconvex methods used in practice, which often succeed without clever initializations.

Recovering low-rank positive semidefinite matrices. The phase retrieval problem has a natural generalization to recovering low-rank positive semidefinite matrices. Consider the problem of recovering an unknown rank-r matrix $\boldsymbol{M} \succeq \mathbf{0}$ in $\mathbb{R}^{n \times n}$ from linear measurement of the form $z_k = \operatorname{tr}(\boldsymbol{A}_k \boldsymbol{M})$ with symmetric $\boldsymbol{A}_k$ for $k = 1, \dots, m$. One can solve the problem by considering the "factorized" version: recovering $\boldsymbol{X} \in \mathbb{R}^{n \times r}$ (up to right invertible transform) from measurements $z_k = \operatorname{tr}(\boldsymbol{X}^* \boldsymbol{A}_k \boldsymbol{X})$. This is a natural generalization of GPR, as one can write the GPR measurements as $y_k^2 = |\boldsymbol{a}_k^* \boldsymbol{x}|^2 = \boldsymbol{x}^* (\boldsymbol{a}_k \boldsymbol{a}_k^*) \boldsymbol{x}$. This generalization and related problems have recently been studied in [SRO15, ZL15, TBSR15, CW15].

12.5 Notations and Wirtinger calculus

Basic notations and facts. Throughout the part, we define complex inner product as: $\langle \boldsymbol{a}, \boldsymbol{b} \rangle \doteq \boldsymbol{a}^* \boldsymbol{b}$ for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{C}^n$. We use $\mathbb{CS}^{n-1}$ for the complex unit sphere in $\mathbb{C}^n$. $\mathbb{CS}^{n-1}(\lambda)$ with $\lambda > 0$ denotes the centered complex sphere with radius λ in $\mathbb{C}^n$. Similarly, we use $\mathbb{CB}^n(\lambda)$ to denote the centered complex ball of radius λ. We use $\mathcal{CN}(k)$ for a standard complex Gaussian vector of length k defined in (12.1.2).

Let $\Re(\boldsymbol{z}) \in \mathbb{R}^n$ and $\Im(\boldsymbol{z}) \in \mathbb{R}^n$ denote the real and imaginary part of a complex vector $\boldsymbol{z} \in \mathbb{C}^n$. It is easy to see that two complex vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are orthogonal in the geometric (real) sense if and only if $\Re(\boldsymbol{w}^* \boldsymbol{z}) = 0$.

For any $\boldsymbol{z}$, obviously $f(\boldsymbol{z}) = f(\boldsymbol{z} e^{i\phi})$ for all ϕ, and the set $\left\{\boldsymbol{z} e^{i\phi} : \phi \in [0, 2\pi)\right\}$ forms a one-dimensional (in the real sense) circle in $\mathbb{C}^n$. Throughout the paper, we reserve $\boldsymbol{x}$ for the unknown target signal, and define the target set as $X \doteq \left\{\boldsymbol{x} e^{i\phi} : \phi \in [0, 2\pi)\right\}$. Moreover, we define

$$\phi(\boldsymbol{z}) \doteq \operatorname*{arg\,min}_{\phi \in [0, 2\pi)} \left\| \boldsymbol{z} - \boldsymbol{x} e^{i\phi} \right\|, \quad \boldsymbol{h}(\boldsymbol{z}) \doteq \boldsymbol{z} - \boldsymbol{x} e^{i\phi(\boldsymbol{z})}, \quad \operatorname{dist}(\boldsymbol{z}, X) \doteq \|\boldsymbol{h}(\boldsymbol{z})\|. \tag{12.5.1}$$

for any $\boldsymbol{z} \in \mathbb{C}^n$. It is not difficult to see that $\Im\left(\boldsymbol{z}^* \boldsymbol{x} e^{i\phi(\boldsymbol{z})}\right) = 0$ and also $\Re\left(\boldsymbol{z}^* \boldsymbol{x} e^{i\phi(\boldsymbol{z})}\right) = |\boldsymbol{x}^* \boldsymbol{z}|$. Moreover, $\boldsymbol{z}_T \doteq i\boldsymbol{z} / \|\boldsymbol{z}\|$ and $-\boldsymbol{z}_T$ are the unit vectors tangent to the circle $\left\{\boldsymbol{z} e^{i\phi} : \phi \in [0, 2\pi)\right\}$ at point $\boldsymbol{z}$.

Wirtinger calculus. Consider a real-valued function $g(\boldsymbol{z}) : \mathbb{C}^n \mapsto \mathbb{R}$. Unless g is constant, it is not complex differentiable. However, if one identifies $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ and treats g as a function in the real domain, g may still be differentiable in the real sense. Doing calculus for g directly in the real domain tends to produce cumbersome expressions. A more elegant way is adopting the Wirtinger calculus, which can be thought of a neat way of organizing the real partial derivatives. Here we only provide a minimal exposition of Wirtinger calculus; similar exposition is also given in [CLS15b]. A systematic development with emphasis on applications in optimization is provided in the article [KD09].

Let $\boldsymbol{z} = \boldsymbol{x} + \mathrm{i}\boldsymbol{y}$ where $\boldsymbol{x} = \Re(\boldsymbol{z})$ and $\boldsymbol{y} = \Im(\boldsymbol{z})$. For a complex-valued function $g(\boldsymbol{z}) = u(\boldsymbol{x}, \boldsymbol{y}) + \mathrm{i}v(\boldsymbol{x}, \boldsymbol{y})$, the Wirtinger derivative is well defined so long as the real-valued functions u and v are differentiable with respect to (w.r.t.) $\boldsymbol{x}$ and $\boldsymbol{y}$. Under these conditions, the Wirtinger derivatives can be defined *formally* as

$$\begin{aligned}\frac{\partial g}{\partial \boldsymbol{z}} &\doteq \left.\frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial \boldsymbol{z}}\right|_{\overline{\boldsymbol{z}}\text{ constant}} = \left.\left[\frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial z_1}, \dots, \frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial z_n}\right]\right|_{\overline{\boldsymbol{z}}\text{ constant}} \\ \frac{\partial g}{\partial \overline{\boldsymbol{z}}} &\doteq \left.\frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial \overline{\boldsymbol{z}}}\right|_{\boldsymbol{z}\text{ constant}} = \left.\left[\frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial \overline{z_1}}, \dots, \frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial \overline{z_n}}\right]\right|_{\boldsymbol{z}\text{ constant}}.\end{aligned}$$

The notation above should only be taken at a formal level. Basically it says when evaluating $\partial g / \partial \boldsymbol{z}$, one just treats $\overline{\boldsymbol{z}}$ as if it was a constant, and vise versa. To evaluate the individual partial derivatives, such as $\frac{\partial g(\boldsymbol{z}, \overline{\boldsymbol{z}})}{\partial z_i}$, all the usual rules of calculus apply. [7]

Note that above the partial derivatives $\frac{\partial g}{\partial \boldsymbol{z}}$ and $\frac{\partial g}{\partial \overline{\boldsymbol{z}}}$ are row vectors. The Wirtinger gradient and Hessian are defined as

$$\nabla g(\boldsymbol{z}) = \left[\frac{\partial g}{\partial \boldsymbol{z}}, \frac{\partial g}{\partial \overline{\boldsymbol{z}}}\right]^* \quad \nabla^2 g(\boldsymbol{z}) = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{z}}\left(\frac{\partial g}{\partial \boldsymbol{z}}\right)^* & \frac{\partial}{\partial \overline{\boldsymbol{z}}}\left(\frac{\partial g}{\partial \boldsymbol{z}}\right)^* \\ \frac{\partial}{\partial \boldsymbol{z}}\left(\frac{\partial g}{\partial \overline{\boldsymbol{z}}}\right)^* & \frac{\partial}{\partial \overline{\boldsymbol{z}}}\left(\frac{\partial g}{\partial \overline{\boldsymbol{z}}}\right)^* \end{bmatrix}, \tag{12.5.2}$$

where we sometimes write $\nabla_{\boldsymbol{z}} g \doteq \left(\frac{\partial g}{\partial \boldsymbol{z}}\right)^*$ and naturally $\nabla_{\overline{\boldsymbol{z}}} g \doteq \left(\frac{\partial g}{\partial \overline{\boldsymbol{z}}}\right)^*$. With gradient and Hessian, the second-order Taylor expansion of $g(\boldsymbol{z})$ at a point $\boldsymbol{z}_0$ is defined as

$$\widehat{g}(\boldsymbol{\delta}; \boldsymbol{z}_0) = g(\boldsymbol{z}_0) + (\nabla g(\boldsymbol{z}_0))^* \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla^2 g(\boldsymbol{z}_0) \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}.$$

For numerical optimization, we are most interested in real-valued g. A real-valued g is stationary at a point $\boldsymbol{z}$ if and only if

$$\nabla_{\boldsymbol{z}} g(\boldsymbol{z}) = \boldsymbol{0}.$$

This is equivalent to the condition $\nabla_{\overline{\boldsymbol{z}}} g = \boldsymbol{0}$, as $\nabla_{\boldsymbol{z}} g = \overline{\nabla_{\overline{\boldsymbol{z}}} g}$ when g is real-valued. The curvature of g at a stationary point $\boldsymbol{z}$ is dictated by the Wirtinger Hessian $\nabla^2 g(\boldsymbol{z})$. An important technical point is that the Hessian quadratic form involves left and right multiplication with a $2n$-dimensional vector consisting of a conjugate pair $(\boldsymbol{\delta}, \overline{\boldsymbol{\delta}})$.

[7] The precise definition is as follows: write $\boldsymbol{z} = \boldsymbol{u} + \mathrm{i}\boldsymbol{v}$. Then $\frac{\partial g}{\partial \boldsymbol{z}} \doteq \frac{1}{2}\left(\frac{\partial g}{\partial \boldsymbol{u}} - \mathrm{i}\frac{\partial g}{\partial \boldsymbol{v}}\right)$. Similarly, $\frac{\partial g}{\partial \overline{\boldsymbol{z}}} \doteq \frac{1}{2}\left(\frac{\partial g}{\partial \boldsymbol{u}} + \mathrm{i}\frac{\partial g}{\partial \boldsymbol{v}}\right)$.

For our particular function $f(\boldsymbol{z}) : \mathbb{C}^n \mapsto \mathbb{R}$ defined in (12.1.1), direct calculation gives

$$\nabla f(\boldsymbol{z}) = \frac{1}{m} \sum_{k=1}^{m} \begin{bmatrix} \left(\left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 - y_k^2\right) \left(\boldsymbol{a}_k \boldsymbol{a}_k^*\right) \boldsymbol{z} \\ \left(\left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 - y_k^2\right) \left(\boldsymbol{a}_k \boldsymbol{a}_k^*\right)^\top \overline{\boldsymbol{z}} \end{bmatrix}, \tag{12.5.3}$$

$$\nabla^2 f(\boldsymbol{z}) = \frac{1}{m} \sum_{k=1}^{m} \begin{bmatrix} \left(2\left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 - y_k^2\right) \boldsymbol{a}_k \boldsymbol{a}_k^* & \left(\boldsymbol{a}_k^* \boldsymbol{z}\right)^2 \boldsymbol{a}_k \boldsymbol{a}_k^\top \\ \left(\boldsymbol{z}^* \boldsymbol{a}_k\right)^2 \overline{\boldsymbol{a}_k} \boldsymbol{a}_k^* & \left(2\left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 - y_k^2\right) \overline{\boldsymbol{a}_k} \boldsymbol{a}_k^\top \end{bmatrix}. \tag{12.5.4}$$

Following the above notation, we write $\nabla_{\boldsymbol{z}} f(\boldsymbol{z})$ and $\nabla_{\overline{\boldsymbol{z}}} f(\boldsymbol{z})$ for denoting the first and second half of $\nabla f(\boldsymbol{z})$, respectively.

Chapter 13

The Geometry of the Objective Function

> The mathematician's patterns, like the painter's or the poet's, must be beautiful; the ideas, like the colors or the words, must fit together in a harmonious way. Beauty is the first test: there is no permanent place in the world for ugly mathematics.
>
> G.F. Hardy

The low-dimensional example described in the introduction (Figure 12.2) provides some clues about the high-dimensional geometry of the objective function $f(\boldsymbol{z})$. Its properties can be seen most clearly through the population objective function $\mathbb{E}_{\boldsymbol{a}}[f(\boldsymbol{z})]$, which can be thought of as a "large sample" version in which $m \to \infty$. We characterize this large-sample geometry in Section 13.1. In Section 13.2, we show that the most important characteristics of this large-sample geometry are present even when the number of observations m is close to the number of degrees of freedom n in the target $\boldsymbol{x}$. Section 13.3 describes several technical problems that arise in the finite sample analysis, and states a number of key intermediate results, which are proved in Chapter 17.

13.1 A Glimpse of the asymptotic function landscape

To characterize the geometry of $\mathbb{E}_{\boldsymbol{a}}[f(\boldsymbol{z})]$ (written as $\mathbb{E}[f]$ henceforth), we simply calculate the expectation of the first and second derivatives of f at each point $\boldsymbol{z} \in \mathbb{C}^n$. We characterize the location of the critical points of the expectation, and use second derivative information to characterize their signatures. An important conclusion is that every local minimum of $\mathbb{E}[f]$ is of the form $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}$, and that all other critical points have a direction of strict negative curvature:

> **Theorem 13.1** *When $\boldsymbol{x} \neq \mathbf{0}$, the only critical points of $\mathbb{E}[f]$ are $\mathbf{0}$, X and $\mathcal{S} \doteq \left\{ \boldsymbol{z} : \boldsymbol{x}^* \boldsymbol{z} = 0,\ \|\boldsymbol{z}\| = \|\boldsymbol{x}\| / \sqrt{2} \right\}$, which are the local maximizer, the set of local/global minimizers, and the set of saddle points, respectively. Moreover, the saddle points and local maximizer have negative curvature in the $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}$ direction.*

Proof We show the statement by partitioning the space $\mathbb{C}^n$ into several regions and analyzing each region individually using the expected gradient and Hessian. These are calculated in Lemma 17.1, and reproduced below:

$$\mathbb{E}[f] = \|\boldsymbol{x}\|^4 + \|\boldsymbol{z}\|^4 - \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 - |\boldsymbol{x}^* \boldsymbol{z}|^2, \tag{13.1.1}$$

$$\nabla \mathbb{E}[f] = \begin{bmatrix} \nabla_{\boldsymbol{z}} \mathbb{E}[f] \\ \nabla_{\overline{\boldsymbol{z}}} \mathbb{E}[f] \end{bmatrix} = \begin{bmatrix} \left(2\|\boldsymbol{z}\|^2 \boldsymbol{I} - \|\boldsymbol{x}\|^2 \boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^* \right) \boldsymbol{z} \\ \left(2\|\boldsymbol{z}\|^2 \boldsymbol{I} - \|\boldsymbol{x}\|^2 \boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^* \right) \overline{\boldsymbol{z}} \end{bmatrix}, \tag{13.1.2}$$

$$\nabla^2 \mathbb{E}[f] = \begin{bmatrix} 2\boldsymbol{z}\boldsymbol{z}^* - \boldsymbol{x}\boldsymbol{x}^* + \left(2\|\boldsymbol{z}\|^2 - \|\boldsymbol{x}\|^2 \right) \boldsymbol{I} & 2\boldsymbol{z}\boldsymbol{z}^\top \\ 2\overline{\boldsymbol{z}}\boldsymbol{z}^* & 2\overline{\boldsymbol{z}}\boldsymbol{z}^\top - \overline{\boldsymbol{x}}\boldsymbol{x}^\top + \left(2\|\boldsymbol{z}\|^2 - \|\boldsymbol{x}\|^2 \right) \boldsymbol{I} \end{bmatrix}. \tag{13.1.3}$$

Based on this, we observe:

- $\boldsymbol{z} = \mathbf{0}$ is a critical point, and the Hessian

  $$\nabla^2 \mathbb{E}[f(\mathbf{0})] = \operatorname{diag}\left(-\boldsymbol{x}\boldsymbol{x}^* - \|\boldsymbol{x}\|^2 \boldsymbol{I}, -\overline{\boldsymbol{x}}\boldsymbol{x}^\top - \|\boldsymbol{x}\|^2 \boldsymbol{I} \right) \prec \mathbf{0}.$$

  Hence, $\boldsymbol{z} = \mathbf{0}$ is a local maximizer.

- In the region $\left\{ \boldsymbol{z} : 0 < \|\boldsymbol{z}\|^2 < \frac{1}{2} \|\boldsymbol{x}\|^2 \right\}$, we have

  $$\begin{bmatrix} \boldsymbol{z} \\ \overline{\boldsymbol{z}} \end{bmatrix}^* \nabla \mathbb{E}[f] = 2\left(2\|\boldsymbol{z}\|^2 - \|\boldsymbol{x}\|^2 \right) \|\boldsymbol{z}\|^2 - 2|\boldsymbol{x}^* \boldsymbol{z}|^2 < 0.$$

  So there is no critical point in this region.

- When $\|\boldsymbol{z}\|^2 = \frac{1}{2}\|\boldsymbol{x}\|^2$, the gradient is $\nabla_{\boldsymbol{z}} \mathbb{E}[f] = -\boldsymbol{x}\boldsymbol{x}^* \boldsymbol{z}$. The gradient vanishes whenever $\boldsymbol{z} \in \operatorname{null}(\boldsymbol{x}\boldsymbol{x}^*)$, which is true if and only if $\boldsymbol{x}^* \boldsymbol{z} = 0$. Thus, we can see that any $\boldsymbol{z} \in \mathcal{S}$ is a critical point. Moreover, for any $\boldsymbol{z} \in \mathcal{S}$,

  $$\begin{bmatrix} \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}} \end{bmatrix}^* \nabla^2 \mathbb{E}[f] \begin{bmatrix} \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}} \end{bmatrix} = -2\|\boldsymbol{x}\|^4.$$

Similarly, one can show that in $\boldsymbol{z}$ direction there is positive curvature. Hence, every $\boldsymbol{z} \in \mathcal{S}$ is a saddle point.

- In the region $\left\{\boldsymbol{z} : \frac{1}{2}\|\boldsymbol{x}\|^2 < \|\boldsymbol{z}\|^2 < \|\boldsymbol{x}\|^2\right\}$, any potential critical point must satisfy

$$\left(2\|\boldsymbol{z}\|^2 - \|\boldsymbol{x}\|^2\right)\boldsymbol{z} = \boldsymbol{x}\boldsymbol{x}^*\boldsymbol{z}.$$

In other words, $2\|\boldsymbol{z}\|^2 - \|\boldsymbol{x}\|^2$ is the positive eigenvalue of the rank-one PSD Hermitian matrix $\boldsymbol{x}\boldsymbol{x}^*$. Hence $2\|\boldsymbol{z}\|^2 - \|\boldsymbol{x}\|^2 = \|\boldsymbol{x}\|^2$. This would imply that $\|\boldsymbol{z}\| = \|\boldsymbol{x}\|$, which does not occur in this region.

- When $\|\boldsymbol{z}\|^2 = \|\boldsymbol{x}\|^2$, critical points must satisfy

$$\left(\|\boldsymbol{x}\|^2 \boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^*\right)\boldsymbol{z} = \boldsymbol{0},$$

and so $\boldsymbol{z} \notin \operatorname{null}(\boldsymbol{x}\boldsymbol{x}^*)$. Given that $\|\boldsymbol{z}\| = \|\boldsymbol{x}\|$, we must have $\boldsymbol{z} = \boldsymbol{x}e^{i\theta}$ for some $\theta \in [0, 2\pi)$. Since f is a nonnegative function, and $f(\boldsymbol{z}) = 0$ for any $\boldsymbol{z} \in X$, X is indeed also the global optimal set.

- For $\|\boldsymbol{z}\| > \|\boldsymbol{x}\|$, since the gradient $\begin{bmatrix} \boldsymbol{z} \\ \overline{\boldsymbol{z}} \end{bmatrix}^* \nabla \mathbb{E}\left[f(\boldsymbol{z})\right] > 0$, there is no critical point present.

Summarizing the above observations completes the proof. ■

This result suggests that the same qualitative properties that we observed for $f(\boldsymbol{z})$ with $\boldsymbol{z} \in \mathbb{R}^2$ also hold for higher-dimensional, complex $\boldsymbol{z}$. The high-dimensional analysis is facilitated by the unitary invariance of the complex normal distribution – the properties of $\mathbb{E}[f]$ at a given point $\boldsymbol{z}$ depend only the norm of $\boldsymbol{z}$ and its inner product with the target vector $\boldsymbol{x}$, i.e., $\boldsymbol{x}^*\boldsymbol{z}$. In the next section, we will show that the important qualitative aspects of this structure are preserved even when m is as small as $Cn\log^3 n$.

13.2 The finite-sample landscape

The following theorem characterizes the geometry of the objective function $f(\boldsymbol{z})$, when the number of samples m is roughly on the order of the number of degrees of freedom (i.e., n) in the vector $\boldsymbol{x}$. The main conclusion is that the space $\mathbb{C}^n$ can be divided into three regions, in which the objective either exhibits negative curvature, strong gradient, or restricted strong convexity.

The result is not surprising in view of the above characterization of the "large-sample" landscape. The intuition is as follows: since the objective function is a sum of independent random variables, when m is sufficiently large, the function values, gradients and Hessians should be uniformly close to their expectations.

Some care is required in making this intuition precise, however. Because the objective function contains fourth powers of Gaussian random variables, it is heavy tailed. Ensuring that f and its derivatives are uniformly close to their expectations requires $m \geq Cn^2$. This would be quite wasteful, since $\boldsymbol{x}$ has only n degrees of freedom.

Fortunately, when $m \geq Cn \, \mathrm{polylog}(n)$, w.h.p. f *still* has benign global geometry, even though its gradient is not uniformly close to its expectation. Perhaps surprisingly, the heavy tailed behavior of f only helps to *prevent* spurious local minimizers – away from the global minimizers and saddle points, the gradient can be sporadically large, but it cannot be sporadically small. This behavior will follow by expressing the decrease of the function along a certain carefully chosen descent direction as a sum of random variables which are heavy tailed, but are also *nonnegative*. Because they are nonnegative, their deviation below their expectation is bounded, and their lower-tail is well-behaved.

Our geometric characterization of the finite-sample objective function reflects these complexities. We prove that there is a partition of $\mathbb{C}^n$ into regions of negative curvature, large gradient, and restricted strong convexity (near the optimizer $\boldsymbol{x}$). The gradient region is further partitioned into two sub-regions, over which different canonical descent directions are studied. Our main geometric result is as follows:

Theorem 13.2 (Main Geometric Results) *There exist positive absolute constants* $c_a, \; c_b, \; c_c$ *and* C*, such that when* $m \geq Cn \log^3 n$*, it holds with probability at least* $1 - c_a \exp\left(-c_b m / \log m\right) - c_c m^{-1}$ *that* $f(\boldsymbol{z})$ *has no spurious local minimizers and the only local/global minimizers are exactly the target set* X*. More precisely, with the same probability,*

$$
\begin{aligned}
\frac{1}{\|\boldsymbol{x}\|^2}\begin{bmatrix}\boldsymbol{x}e^{i\phi(\boldsymbol{z})}\\ \overline{\boldsymbol{x}}e^{-i\phi(\boldsymbol{z})}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z})\begin{bmatrix}\boldsymbol{x}e^{i\phi(\boldsymbol{z})}\\ \overline{\boldsymbol{x}}e^{-i\phi(\boldsymbol{z})}\end{bmatrix} &\leq -\frac{1}{100}\|\boldsymbol{x}\|^2, & \forall\, \boldsymbol{z} \in \mathcal{R}_1, && \textit{(Negative Curvature)}\\
\frac{\boldsymbol{z}^*\nabla_{\boldsymbol{z}} f(\boldsymbol{z})}{\|\boldsymbol{z}\|} &\geq \frac{1}{1000}\|\boldsymbol{x}\|^2\|\boldsymbol{z}\|, & \forall\, \boldsymbol{z} \in \mathcal{R}_2^{\boldsymbol{z}}, && \textit{(Large Gradient)}\\
\frac{\Re\left(\boldsymbol{h}(\boldsymbol{z})^*\nabla_{\boldsymbol{z}} f(\boldsymbol{z})\right)}{\|\boldsymbol{h}(\boldsymbol{z})\|} &\geq \frac{1}{1000}\|\boldsymbol{x}\|^2\|\boldsymbol{z}\|, & \forall\, \boldsymbol{z} \in \mathcal{R}_2^{\boldsymbol{h}}, && \textit{(Large Gradient)}\\
\begin{bmatrix}\boldsymbol{g}(\boldsymbol{z})\\ \overline{\boldsymbol{g}(\boldsymbol{z})}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z})\begin{bmatrix}\boldsymbol{g}(\boldsymbol{z})\\ \overline{\boldsymbol{g}(\boldsymbol{z})}\end{bmatrix} &\geq \frac{1}{4}\|\boldsymbol{x}\|^2, & \forall\, \boldsymbol{z} \in \mathcal{R}_3, && \textit{(Restricted Strong Convexity)}
\end{aligned}
$$

where $\boldsymbol{h}(\boldsymbol{z})$ is defined in (12.5.1), and

$$\boldsymbol{g}(\boldsymbol{z}) \doteq \begin{cases} \boldsymbol{h}(\boldsymbol{z})/\left\|\boldsymbol{h}(\boldsymbol{z})\right\| & \text{if } \operatorname{dist}(\boldsymbol{z}, X) \neq 0, \\ \boldsymbol{h} \in \mathcal{S} \doteq \{\boldsymbol{h} : \Im(\boldsymbol{h}^* \boldsymbol{z}) = 0, \|\boldsymbol{h}\| = 1\} & \text{if } \boldsymbol{z} \in X. \end{cases}$$

Here the regions $\mathcal{R}_1$, $\mathcal{R}_2^{\boldsymbol{z}}$, $\mathcal{R}_2^{\boldsymbol{h}}$ and $\mathcal{R}_3$ cover $\mathbb{C}^n$, and are defined as

$$\mathcal{R}_1 \doteq \left\{\boldsymbol{z} : 8\left|\boldsymbol{x}^* \boldsymbol{z}\right|^2 + \frac{401}{100}\left\|\boldsymbol{x}\right\|^2 \left\|\boldsymbol{z}\right\|^2 \leq \frac{398}{100}\left\|\boldsymbol{x}\right\|^4\right\}, \quad (13.2.1)$$

$$\mathcal{R}_2^{\boldsymbol{z}} \doteq \left\{\boldsymbol{z} : \Re\left(\langle \boldsymbol{z}, \nabla_{\boldsymbol{z}} \mathbb{E}\left[f\right]\rangle\right) \geq \frac{1}{100}\left\|\boldsymbol{z}\right\|^4 + \frac{1}{500}\left\|\boldsymbol{x}\right\|^2 \left\|\boldsymbol{z}\right\|^2\right\}, \quad (13.2.2)$$

$$\begin{aligned} \mathcal{R}_2^{\boldsymbol{h}} \doteq \Big\{\boldsymbol{z} : &\Re\left(\langle \boldsymbol{h}(\boldsymbol{z}), \nabla_{\boldsymbol{z}} \mathbb{E}\left[f\right]\rangle\right) \geq \frac{1}{250}\left\|\boldsymbol{x}\right\|^2 \left\|\boldsymbol{z}\right\| \left\|\boldsymbol{h}(\boldsymbol{z})\right\|, \\ &\frac{11}{20}\left\|\boldsymbol{x}\right\| \leq \left\|\boldsymbol{z}\right\| \leq \left\|\boldsymbol{x}\right\|, \operatorname{dist}(\boldsymbol{z}, X) \geq \frac{\left\|\boldsymbol{x}\right\|}{3}\Big\}, \end{aligned} \quad (13.2.3)$$

$$\mathcal{R}_3 \doteq \left\{\boldsymbol{z} : \operatorname{dist}(\boldsymbol{z}, X) \leq \frac{1}{\sqrt{7}}\left\|\boldsymbol{x}\right\|\right\}. \quad (13.2.4)$$

Proof The quantitative statements are proved sequentially in Proposition 13.3, Proposition 13.4, Proposition 13.5, Proposition 13.6 and Proposition 13.7 in the next section. We next show X are the only local/global minimizers. Obviously local minimizers will not occur in $\mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}} \cup \mathcal{R}_2^{\boldsymbol{h}}$, as at each such point either the gradient is nonzero, or there is a negative curvature direction. So local/global minimizers can occur only in $\mathcal{R}_3$. From (12.5.3), it is easy to check that $\nabla_{\boldsymbol{z}} f(\boldsymbol{x}e^{i\phi}) = \boldsymbol{0}$ and $f(\boldsymbol{x}e^{i\phi}) = 0$ for any $\phi \in [0, 2\pi)$. Since $f(\boldsymbol{z}) \geq 0$ for all $\boldsymbol{z} \in \mathbb{C}^n$, all elements of X are local/global minimizers. To see there is no other critical point in $\mathcal{R}_3$, note that any point $\boldsymbol{z} \in \mathcal{R}_3 \setminus X$ can be written as

$$\boldsymbol{z} = \boldsymbol{x}e^{i\phi(\boldsymbol{z})} + t\boldsymbol{g}, \quad \boldsymbol{g} \doteq \boldsymbol{h}(\boldsymbol{z})/\left\|\boldsymbol{h}(\boldsymbol{z})\right\|, \ t \doteq \operatorname{dist}(\boldsymbol{z}, X).$$

By the restricted strong convexity we have established, and the integral form of Taylor's theorem in Lemma C.2,

$$f(\boldsymbol{z}) = f(\boldsymbol{x}e^{i\phi(\boldsymbol{z})}) + t\begin{bmatrix}\boldsymbol{g} \\ \overline{\boldsymbol{g}}\end{bmatrix}^* \nabla f(\boldsymbol{x}e^{i\phi(\boldsymbol{z})}) + t^2\int_0^1 (1-s)\begin{bmatrix}\boldsymbol{g} \\ \overline{\boldsymbol{g}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{x}e^{i\phi(\boldsymbol{z})} + st\boldsymbol{g})\begin{bmatrix}\boldsymbol{g} \\ \overline{\boldsymbol{g}}\end{bmatrix} ds \geq \frac{1}{8}\left\|\boldsymbol{x}\right\|^2 t^2.$$

similarly, we obtain

$$\begin{aligned} f(\boldsymbol{x}e^{i\phi(\boldsymbol{z})}) = 0 &\geq f(\boldsymbol{z}) - t\begin{bmatrix}\boldsymbol{g} \\ \overline{\boldsymbol{g}}\end{bmatrix}^* \nabla f(\boldsymbol{z}) + t^2\int_0^1 (1-s)\begin{bmatrix}\boldsymbol{g} \\ \overline{\boldsymbol{g}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z} - st\boldsymbol{g})\begin{bmatrix}\boldsymbol{g} \\ \overline{\boldsymbol{g}}\end{bmatrix} ds \\ &\geq f(\boldsymbol{z}) - \begin{bmatrix}\boldsymbol{g} \\ \overline{\boldsymbol{g}}\end{bmatrix}^* \nabla f(\boldsymbol{z}) + \frac{1}{8}\left\|\boldsymbol{x}\right\|^2 t^2. \end{aligned}$$

Summing up the above two inequalities, we obtain

$$t\begin{bmatrix}\boldsymbol{g}\\ \overline{\boldsymbol{g}}\end{bmatrix}^{*}\nabla f(\boldsymbol{z})\geq\frac{1}{4}\left\|\boldsymbol{x}\right\|^{2}t^{2}\Longrightarrow\left\|\nabla f(\boldsymbol{z})\right\|\geq\frac{1}{4\sqrt{2}}\left\|\boldsymbol{x}\right\|^{2}t,$$

as desired. ■

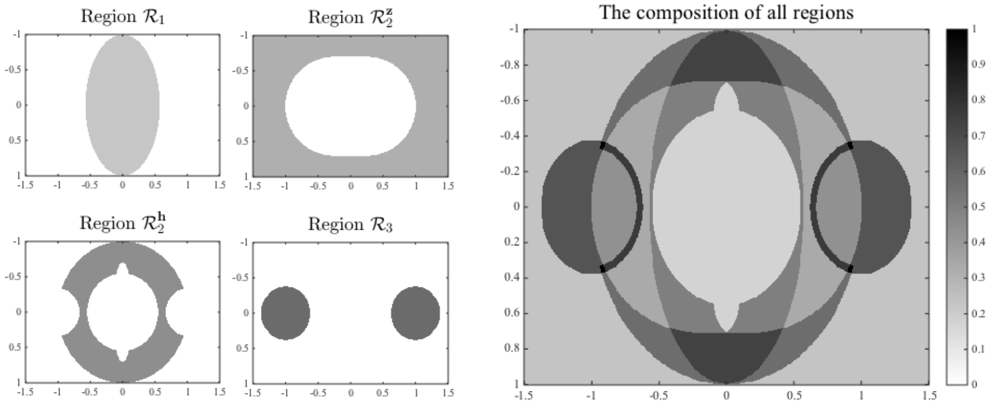


Figure 13.1: Schematic illustration of partitioning regions for Theorem 13.2. This plot corresponds to Figure 12.2, i.e., the target signal is $\boldsymbol{x} = [1; 0]$ and measurements are real Gaussians, such that the function is defined in $\mathbb{R}^2$.

Figure 13.1 visualizes the different regions described in Theorem 13.2, and gives an idea of how they cover the space. For $f(\boldsymbol{z})$, a point $\boldsymbol{z} \in \mathbb{C}^n$ is either near a critical point such that the gradient $\nabla_{\boldsymbol{z}} f(\boldsymbol{z})$ is small (in magnitude), or far from a critical point such that the gradient is large. Any point in $\mathcal{R}_2^{\boldsymbol{z}} \cup \mathcal{R}_2^{\boldsymbol{h}}$ is far from a critical point, as the following is true:

$$\left\|\nabla_{\boldsymbol{z}} f(\boldsymbol{z})\right\| \geq \frac{\boldsymbol{z}^{*}\nabla_{\boldsymbol{z}} f(\boldsymbol{z})}{\left\|\boldsymbol{z}\right\|} \geq \frac{1}{1000}\left\|\boldsymbol{x}\right\|^{2}\left\|\boldsymbol{z}\right\|, \text{ or } \left\|\nabla_{\boldsymbol{z}} f(\boldsymbol{z})\right\| \geq \frac{\Re\left(\boldsymbol{h}(\boldsymbol{z})^{*}\nabla_{\boldsymbol{z}} f(\boldsymbol{z})\right)}{\left\|\boldsymbol{h}(\boldsymbol{z})\right\|} \geq \frac{1}{1000}\left\|\boldsymbol{x}\right\|^{2}\left\|\boldsymbol{z}\right\|.$$

The rest of the space consists of points near critical points. Since $\mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}} \cup \mathcal{R}_2^{\boldsymbol{h}} \cup \mathcal{R}_3$ cover the space, the rest points are included in $\mathcal{R}_1 \cup \mathcal{R}_3$. For any $\boldsymbol{z}$ in $\mathcal{R}_1$, the quantity

$$\frac{1}{\left\|\boldsymbol{x}\right\|^{2}}\begin{bmatrix}\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}\\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})}\end{bmatrix}^{*}\nabla^{2}f(\boldsymbol{z})\begin{bmatrix}\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}\\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})}\end{bmatrix}$$

measures the local curvature of $f(\boldsymbol{z})$ in the $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}$ direction. Strict negativity of this quantity implies that the neighboring critical point is either a local maximizer, or a saddle point. Moreover, $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}$ is a local descent direction, even if $\nabla_{\boldsymbol{z}} f(\boldsymbol{z}) = \mathbf{0}$. For any $\boldsymbol{z} \in \mathcal{R}_3$, $\boldsymbol{g}(\boldsymbol{z})$ is the unit vector that points to $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}$, and is also geometrically orthogonal to the $\mathrm{i}\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}$ which is tangent the circle X at $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}$. The strict positivity of the quantity

$$\begin{bmatrix} \boldsymbol{g}(\boldsymbol{z}) \\ \overline{\boldsymbol{g}(\boldsymbol{z})} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix} \boldsymbol{g}(\boldsymbol{z}) \\ \overline{\boldsymbol{g}(\boldsymbol{z})} \end{bmatrix}$$

implies that locally $f(\boldsymbol{z})$ is strongly convex in $\boldsymbol{g}(\boldsymbol{z})$ direction, although it is flat on the complex circle $\left\{\boldsymbol{z}e^{i\phi} : \phi \in [0, 2\pi)\right\}$. In particular, the result applied to $\boldsymbol{z} \in X$ implies that on X, $f(\boldsymbol{z})$ is strongly convex in any direction orthogonal to X. This observation, together with the fact that the Hessian is Lipschitz, implies that there is a neighborhood of X on which $\boldsymbol{v}^* \nabla^2 f(\boldsymbol{x}) \boldsymbol{v} > 0$ for *every* direction that is orthogonal to the trivial direction $i\boldsymbol{z}$, not just the particular direction $\boldsymbol{g}(\boldsymbol{z})$. This stronger property can be used to study the asymptotic convergence rate of algorithms; in particular, we will use it to obtain quadratic convergence for a certain variant of the trust-region method.

In the asymptotic version, we characterized only the critical points. In this finite-sample version, we characterize the whole space and particularly provide quantitative control for regions near critical points (i.e., $\mathcal{R}_1 \cup \mathcal{R}_3$). These concrete quantities are important for algorithm design and analysis (see Chapter 14).

In sum, our objective $f(\boldsymbol{z})$ has the benign geometry that all local minimizers are global, and each $\boldsymbol{z} \in \mathbb{C}^n$ has either large gradient or directional negative curvature, or lies in the vicinity of a local minimizer around which the function is locally restrictedly strongly convex. Functions with this property lie in the $\mathcal{X}$ family we defined in Chapter 2. As discussed therein, functions in this class admit simple iterative methods (including the noisy gradient method, curvilinear search, and trust-region methods), which avoid being trapped near saddle points, and efficiently obtain a global minimizer.

13.3 Key steps in the geometric analysis

Our proof strategy is fairly simple: we work out uniform bounds on the quantities for each of the four regions, and finally show the regions together cover the space. Since (12.1.1) and associated derivatives take the form of summation of m independent random variables, the proof involves concentration and covering arguments [Ver12]. The main challenge in our argument will be the heavy tailed nature of f and its gradient.

Proposition 13.3 *When $m \geq Cn \log n$, it holds with probability at least $1 - c_a \exp\left(-c_b m / \log m\right) - c_c m^{-1}$*

that

$$\frac{1}{\|\boldsymbol{x}\|^2} \begin{bmatrix} \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix} \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})} \end{bmatrix} \le -\frac{1}{100}\|\boldsymbol{x}\|^2$$

for all $\boldsymbol{z} \in \mathcal{R}_1$ defined in (13.2.1). Here C, and c_a to c_c are positive absolute constants.

Proof See Section 17.2 on Page 169. ■

The expected gradient $\nabla_{\boldsymbol{z}}\mathbb{E}[f(\boldsymbol{z})]$ is a linear combination of $\boldsymbol{z}$ and $\boldsymbol{x}$. We will divide $\mathcal{R}_2$ into two overlapped regions, $\mathcal{R}_2^{\boldsymbol{z}}$ and $\mathcal{R}_2^{\boldsymbol{h}}$, roughly matching the case

$$\Re\left(\boldsymbol{z}^*\nabla_{\boldsymbol{z}}\mathbb{E}[f(\boldsymbol{z})]\right) > 0$$

and the case

$$\Re\left(\left(\boldsymbol{z} - \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}\right)^* \nabla_{\boldsymbol{z}}\mathbb{E}[f(\boldsymbol{z})]\right) > 0,$$

respectively.

Proposition 13.4 *When $m \ge Cn\log n$, it holds with probability at least $1 - c_a\exp(-c_b m/\log m) - c_c m^{-1}$ that*

$$\frac{\boldsymbol{z}^*\nabla_{\boldsymbol{z}} f(\boldsymbol{z})}{\|\boldsymbol{z}\|} \ge \frac{1}{1000}\|\boldsymbol{x}\|^2\|\boldsymbol{z}\|$$

for all $\boldsymbol{z} \in \mathcal{R}_2^{\boldsymbol{z}}$ defined in (13.2.2). Here C and c_a to c_c are positive absolute constants.

Proof See Section 17.3 on Page 170. ■

Proposition 13.5 *When $m \ge Cn\log^3 n$, it holds with probability at least $1 - c_a\exp(-c_b m/\log^2 m) - c_c m^{-1}$ that*

$$\Re\left(\boldsymbol{h}(\boldsymbol{z})^*\nabla_{\boldsymbol{z}} f(\boldsymbol{z})\right) \ge \frac{1}{1000}\|\boldsymbol{x}\|^2\|\boldsymbol{z}\|\|\boldsymbol{h}(\boldsymbol{z})\|$$

for all $\boldsymbol{z} \in \mathcal{R}_2^{\boldsymbol{h}}$ defined in (13.2.3). Here c_a to c_c and C are positive absolute constants.

Proof See Section 17.4 on Page 171. ■

Next, we show that for any $\boldsymbol{z} \in \mathbb{C}^n$ near X, the objective f is strongly convex in the direction $\boldsymbol{z} - \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}$. This allows us to achieve a quadratic asymptotic rate of convergence with the modified trust-region algorithm we propose later.

Proposition 13.6 *When $m \geq Cn \log n$ for a sufficiently large constant C, it holds with probability at least $1 - c_a m^{-1} - c_b \exp(-c_c m / \log m)$ that*

$$\begin{bmatrix} \boldsymbol{g}(\boldsymbol{z}) \\ \overline{\boldsymbol{g}(\boldsymbol{z})} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix} \boldsymbol{g}(\boldsymbol{z}) \\ \overline{\boldsymbol{g}(\boldsymbol{z})} \end{bmatrix} \geq \frac{1}{4} \|\boldsymbol{x}\|^2$$

for all $\boldsymbol{z} \in \mathcal{R}_3$ defined in (13.2.4) and for all

$$\boldsymbol{g}(\boldsymbol{z}) \doteq \begin{cases} \left(\boldsymbol{z} - \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}\right) / \left\|\boldsymbol{z} - \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}\right\| & \text{if } \operatorname{dist}(\boldsymbol{z}, X) \neq 0, \\ \boldsymbol{h} \in \mathcal{S} \doteq \{\boldsymbol{h} : \Im(\boldsymbol{h}^* \boldsymbol{z}) = 0, \|\boldsymbol{h}\| = 1\} & \text{if } \boldsymbol{z} \in X. \end{cases}$$

Here C, c_a to c_c are positive absolute constants.

Proof See Section 17.5 on Page 174. ■

Finally, we show that the regions we defined above cover the whole space. Formally,

Proposition 13.7 *We have $\mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}} \cup \mathcal{R}_2^{\boldsymbol{h}} \cup \mathcal{R}_3 = \mathbb{C}^n$.*

Proof See Section 17.6 on Page 175. ■

The main challenge is that the function (12.1.1) is fourth-order polynomial, and most quantities arising in the above propositions involve heavy-tailed random variables. For example, we need to control

$$\frac{1}{m} \sum_{k=1}^m |\boldsymbol{a}_k^* \boldsymbol{z}|^4 \quad \text{for all } \boldsymbol{z} \in \mathcal{R}_2^{\boldsymbol{z}} \tag{13.3.1}$$

in proving Proposition 13.4,

$$\frac{1}{m} \sum_{k=1}^m |\boldsymbol{a}_k^* \boldsymbol{z}|^2 \, \Re\left((\boldsymbol{z} - \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^* \boldsymbol{a}_k \boldsymbol{a}_k^* \boldsymbol{z}\right) \quad \text{for all } \boldsymbol{z} \in \mathcal{R}_2^{\boldsymbol{h}} \tag{13.3.2}$$

in proving Proposition 13.5, and a quantity of the form

$$\frac{1}{m} \sum_{k=1}^m |\boldsymbol{a}_k^* \boldsymbol{w}|^2 \, |\boldsymbol{a}_k^* \boldsymbol{z}|^2 \quad \text{for all } \boldsymbol{w}, \boldsymbol{z} \tag{13.3.3}$$

in proving Proposition 13.6. With only $Cn \log^3 n$ samples, these quantities do not concentrate uniformly about their expectations. Fortunately, this heavy-tailed behavior does not prevent the objective function from being globally well-structured for optimization. Our bounds on the gradient and Hessian depend only on the *lower tails* of the above quantities. For (13.3.1) and (13.3.3) that are sum of independent nonnegative random variables, the lower tails concentrate uniformly as these lower-bounded variables are sub-Gaussian viewed from lower tails (see Lemma A.6 and Lemma 17.4). For (13.3.2), we carefully construct a proxy quantity that is summation of bounded random variables which uniformly bounds (13.3.2) from below.

Chapter 14

Optimization by Trust-Region Method (TRM)

> The purpose of computing is insight, not numbers.
>
> Richard Hamming

Based on the geometric characterization in the preceding chapter, we describe a second-order trust-region algorithm that produces a close approximation (i.e., up to numerical precision) to the global minimizer of (12.1.1) in polynomial number of steps. One interesting aspect of f in the complex space is that each point has a "circle" of equivalent points that have the same function value. Thus, we constrain each step to move "orthogonal" to the trivial direction. This simple modification helps the algorithm to converge faster in practice, and proves important to the quadratic asymptotic convergence rate in theory.

14.1 A modified trust-region algorithm

The basic idea of the trust-region method is simple: we generate a sequence of iterates $\boldsymbol{z}^{(0)}, \boldsymbol{z}^{(1)}, \dots$, by repeatedly constructing quadratic approximations $\widehat{f}(\boldsymbol{\delta}; \boldsymbol{z}^{(r)}) \approx f(\boldsymbol{z}^{(r)} + \boldsymbol{\delta})$, minimizing $\widehat{f}$ to obtain a step $\boldsymbol{\delta}$, and setting $\boldsymbol{z}^{(r+1)} = \boldsymbol{z}^{(r)} + \boldsymbol{\delta}$. More precisely, we approximate $f(\boldsymbol{z})$ around $\boldsymbol{z}^{(r)}$ using the second-order Taylor expansion,

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{z}^{(r)}) = f(\boldsymbol{z}^{(r)}) + \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r)}) + \frac{1}{2} \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}^{(r)}) \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix},$$

and solve

$$\operatorname{minimize}_{\boldsymbol{\delta}\in\mathbb{C}^n}\ \widehat{f}(\boldsymbol{\delta};\boldsymbol{z}^{(r)}),\quad \text{subject to}\quad \Im\left(\boldsymbol{\delta}^*\boldsymbol{z}^{(r)}\right)=0,\quad \|\boldsymbol{\delta}\|\le\Delta, \tag{14.1.1}$$

to obtain the step $\boldsymbol{\delta}$. In (14.1.1), Δ controls the trust-region size. The first linear constraint further forces the movement $\boldsymbol{\delta}$ to be geometrically orthogonal to the $\mathrm{i}\boldsymbol{z}$ direction, along which the possibility for reducing the function value is limited. Enforcing this linear constraint is a strategic modification to the classical trust-region subproblem.

Reduction to the standard trust-region subproblem. The modified trust-region subproblem is easily seen to be equivalent to a classical trust-region subproblem (with no constraint) over $2n-1$ real variables. Notice that $\left\{\boldsymbol{w}\in\mathbb{C}^n:\Im(\boldsymbol{w}^*\boldsymbol{z}^{(r)})=0\right\}$ forms a subspace of dimension $2n-1$ over $\mathbb{R}^{2n}$. Take any matrix $\boldsymbol{U}(\boldsymbol{z}^{(r)})\in\mathbb{C}^{n\times(2n-1)}$ whose columns form an orthonormal basis for the subspace, i.e., $\Re(\boldsymbol{U}_i^*\boldsymbol{U}_j)=\delta_{ij}$ for any columns $\boldsymbol{U}_i$ and $\boldsymbol{U}_j$. The subproblem can then be reformulated as ($\boldsymbol{U}$ short for $\boldsymbol{U}(\boldsymbol{z}^{(r)})$)

$$\operatorname{minimize}_{\boldsymbol{\xi}\in\mathbb{R}^{2n-1}}\ \widehat{f}(\boldsymbol{U}\boldsymbol{\xi};\boldsymbol{z}^{(r)}),\quad \text{subject to}\quad \|\boldsymbol{\xi}\|\le\Delta. \tag{14.1.2}$$

Let us define

$$\boldsymbol{g}(\boldsymbol{z}^{(r)})\doteq\begin{bmatrix}\boldsymbol{U}\\ \overline{\boldsymbol{U}}\end{bmatrix}^*\nabla f(\boldsymbol{z}^{(r)}),\quad \boldsymbol{H}(\boldsymbol{z}^{(r)})\doteq\begin{bmatrix}\boldsymbol{U}\\ \overline{\boldsymbol{U}}\end{bmatrix}^*\nabla^2 f(\boldsymbol{z}^{(r)})\begin{bmatrix}\boldsymbol{U}\\ \overline{\boldsymbol{U}}\end{bmatrix}. \tag{14.1.3}$$

Then, the quadratic approximation of $f(\boldsymbol{z})$ around $\boldsymbol{z}^{(r)}$ can be rewritten as

$$\widehat{f}(\boldsymbol{\xi};\boldsymbol{z}^{(r)})=f(\boldsymbol{z}^{(r)})+\boldsymbol{\xi}^\top\boldsymbol{g}(\boldsymbol{z}^{(r)})+\frac{1}{2}\boldsymbol{\xi}^\top\boldsymbol{H}(\boldsymbol{z}^{(r)})\boldsymbol{\xi}. \tag{14.1.4}$$

By structure of the Wirtinger gradient $\nabla f(\boldsymbol{z}^{(r)})$ and Wirtinger Hessian $\nabla^2 f(\boldsymbol{z}^{(r)})$, $\boldsymbol{g}(\boldsymbol{z}^{(r)})$ and $\boldsymbol{H}(\boldsymbol{z}^{(r)})$ contain only real entries.

So, any method which can solve the classical trust-region subproblem can be directly applied to the modified problem (14.1.1). Although the resulting problem can be nonconvex, it can be solved in polynomial time, by root-finding or SDP relaxations. Our convergence guarantees assume an exact solution of this problem; we outline how to obtain such a solution via SDP relaxation. In practice, though, even very inexact solutions of the trust-region subproblem suffice.[1] Inexact iterative solvers for the trust-region subproblem can be engineered to avoid the need to densely represent the Hessian; these methods have the attractive

[1] This can also be proved, in a relatively straightforward way, using the geometry of the objective f. In the interest of brevity, we do not pursue this here.

property that they attempt to optimize the amount of Hessian information that is used at each iteration, in order to balance rate of convergence and computation.

In the interest of theory, we describe briefly how to apply SDP relaxation to solve problem (14.1.2). This SDP relaxation has the important property that it is always exact, even if the Hessian is indefinite. By introducing

$$\widehat{\boldsymbol{\xi}} = \begin{bmatrix} \boldsymbol{\xi} \\ 1 \end{bmatrix}, \quad \boldsymbol{\Xi} = \widehat{\boldsymbol{\xi}}\, \widehat{\boldsymbol{\xi}}^\top, \quad \boldsymbol{M} = \begin{bmatrix} \boldsymbol{H}(\boldsymbol{z}^{(r)}) & \boldsymbol{g}(\boldsymbol{z}^{(r)}) \\ \boldsymbol{g}(\boldsymbol{z}^{(r)})^\top & 0 \end{bmatrix},$$

we can lift problem (14.1.2) as a semidefinite program (SDP):

$$\min_{\boldsymbol{\Xi}} \langle \boldsymbol{\Xi}, \boldsymbol{M} \rangle, \quad \text{s.t.} \quad \mathrm{tr}\,(\boldsymbol{\Xi}) \le \Delta^2 + 1, \quad \langle \boldsymbol{E}_{2n}, \boldsymbol{\Xi} \rangle = 1, \quad \boldsymbol{\Xi} \succeq \boldsymbol{0}, \tag{14.1.5}$$

where $\boldsymbol{E}_{2n} = \boldsymbol{e}_{2n}\boldsymbol{e}_{2n}^\top$, and $\langle \cdot, \cdot \rangle$ reduces to the usual real inner product of real-valued matrices.

Once the subproblem (14.1.5) is solved to optimal $\boldsymbol{\Xi}_\star$, we can perform eigen-decomposition on $\boldsymbol{\Xi}_\star$ as $\boldsymbol{\Xi}_\star = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. Let $\boldsymbol{v}$ be the principle eigenvector of $\boldsymbol{V}$, and let $\boldsymbol{\xi}_\star$ be the first $2n-1$ coordinate of $\boldsymbol{v}$, then the optimum of the original TRM subproblem (14.1.1) is recovered as $\boldsymbol{\delta}_\star = \boldsymbol{U}\boldsymbol{\xi}_\star$.

14.2 Convergence analysis

Our convergence proof proceeds as follows. Let $\boldsymbol{\delta}^\star$ denote the optimizer of the trust-region subproblem at a point $\boldsymbol{z}$. If $\|\nabla f(\boldsymbol{z})\|$ is bounded away from zero, or $\lambda_{\min}(\nabla^2 f(\boldsymbol{z}))$ is bounded below zero, we can guarantee that that $\widehat{f}(\boldsymbol{\delta}^\star, \boldsymbol{z}) - f(\boldsymbol{z}) < -\varepsilon$, for some ε which depends on our bounds on these quantities. Because $f(\boldsymbol{z} + \boldsymbol{\delta}^\star) \approx \widehat{f}(\boldsymbol{\delta}^\star, \boldsymbol{z}) < f(\boldsymbol{z}) - \varepsilon$, we can guarantee (roughly) an ε decrease in the objective function at each iteration. Because this ε is uniformly bounded away from zero over the gradient and negative curvature regions, the algorithm can take at most finitely many steps in these regions. Once it enters the strong convexity region around the global minimizers, the algorithm behaves much like a typical Newton-style algorithm; in particular, it exhibits asymptotic quadratic convergence. Below, we prove quantitative versions of these statements. We begin by stating several basic facts that are useful for the convergence proof.

14.2.1 Preliminaries

Norm of the target vector and initialization. In our problem formulation, $\|\boldsymbol{x}\|$ is not known ahead of time. However, it can be well estimated. When $\boldsymbol{a} \sim \mathcal{CN}(n)$, $\mathbb{E}\,|\boldsymbol{a}^*\boldsymbol{x}|^2 = \|\boldsymbol{x}\|^2$. By Bernstein's inequality,

$\frac{1}{m}\sum_{k=1}^{m}|\boldsymbol{a}_k^*\boldsymbol{x}|^2 \geq \frac{1}{9}\|\boldsymbol{x}\|^2$ with probability at least $1-\exp(-cm)$. Thus, with the same probability, the quantity $R_0 \doteq 3(\frac{1}{m}\sum_{k=1}^{m}|\boldsymbol{a}_k^*\boldsymbol{x}|^2)^{1/2}$ is an upper bound for $\|\boldsymbol{x}\|$. For the sake of analysis, we will assume the initialization $\boldsymbol{z}^{(0)}$ is an *arbitrary point* over $\mathbb{CB}^n(R_0)$. Now consider a fixed $R_1 > R_0$. By Lemma 17.3, Lemma 17.4, and the fact that $\max_{k\in[m]}\|\boldsymbol{a}_k\|^4 \leq 10n^2\log^2 m$ with probability at least $1-c_a m^{-n}$, we have that the following estimate

$$
\begin{aligned}
&\inf_{\boldsymbol{z},\boldsymbol{z}':\ \|\boldsymbol{z}\|\leq R_0,\ \|\boldsymbol{z}'\|\geq R_1} f(\boldsymbol{z}') - f(\boldsymbol{z}) \\
&= \inf_{\boldsymbol{z},\boldsymbol{z}':\ \|\boldsymbol{z}\|\leq R_0,\ \|\boldsymbol{z}'\|\geq R_1} \frac{1}{m}\sum_{k=1}^{m}\left[|\boldsymbol{a}_k^*\boldsymbol{z}'|^4 - |\boldsymbol{a}_k^*\boldsymbol{z}|^4 - 2\,|\boldsymbol{a}_k^*\boldsymbol{z}'|^2\,|\boldsymbol{a}_k^*\boldsymbol{x}|^2 + 2\,|\boldsymbol{a}_k^*\boldsymbol{z}|^2\,|\boldsymbol{a}_k^*\boldsymbol{x}|^2\right] \\
&\geq \inf_{\boldsymbol{z},\boldsymbol{z}':\ \|\boldsymbol{z}\|\leq R_0,\ \|\boldsymbol{z}'\|\geq R_1} \frac{199}{200}\|\boldsymbol{z}'\|^4 - 10n^2\log^2 m\,\|\boldsymbol{z}\|^4 - \frac{201}{200}\left(\|\boldsymbol{z}'\|^2\|\boldsymbol{x}\|^2 + |\boldsymbol{x}^*\boldsymbol{z}'|^2\right) \\
&\geq \inf_{\boldsymbol{z}':\|\boldsymbol{z}'\|\geq R_1} \frac{199}{200}\|\boldsymbol{z}'\|^4 - 10n^2\log^2 mR_0^4 - \frac{201}{100}\|\boldsymbol{z}'\|^2 R_0^2
\end{aligned}
$$

holds with probability at least $1-c_b m^{-1} - c_c\exp(-c_d m/\log m)$, provided $m \geq Cn\log n$ for a sufficiently large C. It can be checked that when

$$
R_1 = 3\sqrt{n\log m}R_0, \tag{14.2.1}
$$

we have

$$
\inf_{\boldsymbol{z}':\|\boldsymbol{z}'\|\geq R_1} \frac{199}{200}\|\boldsymbol{z}'\|^4 - 10n^2\log^2 mR_0^4 - \frac{201}{100}\|\boldsymbol{z}'\|^2 R_0^2 \geq 40n^2\log^2 mR_0^4.
$$

Thus, we conclude that when $m \geq Cn\log n$, w.h.p., the sublevel set $\left\{\boldsymbol{z} : f(\boldsymbol{z}) \leq f(\boldsymbol{z}^{(0)})\right\}$ is contained in the set

$$
\Gamma \doteq \mathbb{CB}^n(R_1). \tag{14.2.2}
$$

Lipschitz Properties We write $\boldsymbol{A} \doteq [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m]$ so that $\|\boldsymbol{A}\|_{\ell^1\to\ell^2} = \max_{k\in[m]}\|\boldsymbol{a}_k\|$. We next provide estimates of Lipschitz constants of f and its derivatives, restricted to a slightly larger region than Γ:

> **Lemma 14.1 (Local Lipschitz Properties)** *The Lipschitz constants for $f(\boldsymbol{z})$, $\nabla f(\boldsymbol{z})$, and $\nabla^2 f(\boldsymbol{z})$ over the set $\Gamma' \doteq \mathbb{CB}^n(2R_1)$ can be bounded above by L_f, L_g, and L_h respectively, where*
>
> $$
> L_f \doteq 7\times 10^6\cdot(n\log m)^{\frac{3}{2}}\|\boldsymbol{A}\|_{\ell^1\to\ell^2}^2\|\boldsymbol{x}\|^3, \quad L_g \doteq 19000\sqrt{2}n\log m\|\boldsymbol{A}\|_{\ell^1\to\ell^2}^2\|\boldsymbol{x}\|^2,
> $$
> $$
> L_h \doteq 480\cdot(n\log m)^{\frac{1}{2}}\|\boldsymbol{A}\|_{\ell^1\to\ell^2}^2\|\boldsymbol{x}\|
> $$

with probability at least $1 - c_a \exp(-c_b m)$, provided $m \geq Cn$ for a sufficiently large absolute constant C. Here c_a through c_e are positive absolute constants.

Proof See Section 18.2 on Page 179. ■

Property of Hessians near the Target Set X. Define a region

$$\mathcal{R}_3' \doteq \left\{ \boldsymbol{z} : \|\boldsymbol{h}(\boldsymbol{z})\| \leq \frac{1}{10 L_h} \|\boldsymbol{x}\|^2 \right\}. \tag{14.2.3}$$

We will provide spectral upper and lower bounds for the (restricted) Hessian matrices $\boldsymbol{H}(\boldsymbol{z})$, where $\boldsymbol{H}(\boldsymbol{z})$ is as defined in (14.1.3). These bounds follow by bounding $\boldsymbol{H}(\boldsymbol{z})$ on X, and then using the Lipschitz property of the Hessian to extend the bounds to a slightly larger region around X.

Lemma 14.2 (Lower and Upper Bounds of Restricted Hessian in $\mathcal{R}_3'$) *When $m \geq Cn \log n$, it holds with probability at least $1 - c_a m^{-1} - c_b \exp\left(-c_c m / \log m\right)$ that*

$$m_H \boldsymbol{I} \preceq \boldsymbol{H}(\boldsymbol{z}) \preceq M_H \boldsymbol{I}$$

for all $\boldsymbol{z} \in \mathcal{R}_3'$ with $m_H = 22/25 \|\boldsymbol{x}\|^2$ and $M_H = 9/2 \|\boldsymbol{x}\|^2$. Here C, c_a to c_c are positive absolute constants.

Proof See Section 18.3 on Page 180. ■

14.2.2 Convergence of TRM

We are now ready to prove the convergence of the TRM. Throughout, we will assume $m \geq Cn \log^3 n$ for a sufficiently large constant C, so that all the events of interest hold w.h.p..

Our initialization is an arbitrary point $\boldsymbol{z}^{(0)} \in \mathbb{CB}^n(R_0) \subseteq \Gamma$. We will analyze effect of a trust-region step from any iterate $\boldsymbol{z}^{(r)} \in \Gamma$. Based on these arguments, we will show that whenever $\boldsymbol{z}^{(r)} \in \Gamma$, $\boldsymbol{z}^{(r+1)} \in \Gamma$, and so the entire iterate sequence remains in Γ. The analysis will use the fact that f and its derivatives are Lipschitz over the trust-region $\boldsymbol{z} + \mathbb{CB}^n(\Delta)$. This follows from Proposition 14.1, provided

$$\Delta \leq R_1. \tag{14.2.4}$$

The next auxiliary lemma makes precise the intuition that whenever there exists a descent direction, the step size parameter Δ is sufficiently small, a trust-region step will decrease the objective.

Lemma 14.3 *For any $\boldsymbol{z} \in \Gamma$, suppose there exists a vector $\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\| \leq \Delta$ such that*

$$\Im(\boldsymbol{\delta}^* \boldsymbol{z}) = 0 \quad \textit{and} \quad f(\boldsymbol{z} + \boldsymbol{\delta}) \leq f(\boldsymbol{z}) - d,$$

for a certain $d > 0$. Then the trust-region subproblem (14.1.1) returns a point $\boldsymbol{\delta}_\star$ with $\|\boldsymbol{\delta}_\star\| \leq \Delta$ and

$$f(\boldsymbol{z} + \boldsymbol{\delta}_\star) \leq f(\boldsymbol{z}) - d + \frac{2}{3} L_h \Delta^3.$$

Proof See Section 18.4 on Page 182. ■

The next proposition says when Δ is chosen properly, a trust-region step from a point with negative local curvature decreases the function value by a concrete amount.

Proposition 14.4 (Function Value Decrease in Negative Curvature Region $\mathcal{R}_1$) *Suppose the current iterate $\boldsymbol{z}^{(r)} \in \mathcal{R}_1 \cap \Gamma$, and our trust-region size satisfies*

$$\Delta \leq \frac{1}{400 L_h} \|\boldsymbol{x}\|^2 . \tag{14.2.5}$$

Then an optimizer $\boldsymbol{\delta}_\star$ to (14.1.1) leads to $\boldsymbol{z}^{(r+1)} = \boldsymbol{z}^{(r)} + \boldsymbol{\delta}_\star$ that obeys

$$f(\boldsymbol{z}^{(r+1)}) - f(\boldsymbol{z}^{(r)}) \leq -d_1 \doteq -\frac{1}{400} \Delta^2 \|\boldsymbol{x}\|^2 . \tag{14.2.6}$$

Proof See Section 18.5 on Page 182. ■

The next proposition shows that when Δ is chosen properly, a trust-region step from a point with strong gradient decreases the objective by a concrete amount.

Proposition 14.5 (Function Value Decrease in Large Gradient Region $\mathcal{R}_2$) *Suppose our current iterate $\boldsymbol{z}^{(r)} \in (\mathcal{R}_2^{\boldsymbol{z}} \cup \mathcal{R}_2^{\boldsymbol{h}}) \cap \mathcal{R}_1^c \cap \Gamma$, and our trust-region size satisfies*

$$\Delta \leq \min \left\{ \frac{\|\boldsymbol{x}\|^3}{8000 L_g}, \sqrt{\frac{3 \|\boldsymbol{x}\|^3}{16000 L_h}} \right\} . \tag{14.2.7}$$

Then an optimizer $\boldsymbol{\delta}_\star$ to (14.1.1) leads to $\boldsymbol{z}^{(r+1)} = \boldsymbol{z}^{(r)} + \boldsymbol{\delta}_\star$ that obeys

$$f(\boldsymbol{z}^{(r+1)}) - f(\boldsymbol{z}^{(r)}) \leq -d_2 \doteq -\frac{1}{4000} \Delta \|\boldsymbol{x}\|^3 . \tag{14.2.8}$$

Proof See Section 18.6 on Page 183. ■

Now, we argue about $\mathcal{R}_3$, in which the behavior of the algorithm is more complicated. For the region $\mathcal{R}_3 \setminus \mathcal{R}_3'$, the restricted strong convexity in radial directions around X as established in Proposition 13.6

implies that the gradient at any point in $\mathcal{R}_3 \setminus \mathcal{R}_3'$ is nonzero. Thus, one can treat this as another strong gradient region, and carry out essentially the same argument as in Proposition 14.5.

Proposition 14.6 (Function Value Decrease in $\mathcal{R}_3 \setminus \mathcal{R}_3'$) *Suppose our current iterate $\boldsymbol{z}^{(r)} \in \mathcal{R}_3 \setminus \mathcal{R}_3'$, and our trust-region size satisfies*

$$\Delta \le \min\left\{ \frac{\|\boldsymbol{x}\|^4}{160 L_h L_g}, \sqrt{\frac{3}{320}} \frac{\|\boldsymbol{x}\|^2}{L_h} \right\}. \tag{14.2.9}$$

Then an optimizer $\boldsymbol{\delta}_\star$ to (14.1.1) leads to $\boldsymbol{z}^{(r+1)} = \boldsymbol{z}^{(r)} + \boldsymbol{\delta}_\star$ that obeys

$$f(\boldsymbol{z}^{(r+1)}) - f(\boldsymbol{z}^{(r)}) \le -d_3 \doteq -\frac{1}{80 L_h} \Delta \|\boldsymbol{x}\|^4 . \tag{14.2.10}$$

Proof See Section 18.7 on Page 184. ■

Our next several propositions show that when the iterate sequence finally moves into $\mathcal{R}_3'$, it can be divided into two ordered phases, either of which can be absent: first, constrained steps in which the constraint of the trust-region subproblem is active, and second, unconstrained steps in which the trust-region constraint is inactive. The next proposition shows that when Δ is chosen properly, a constrained step in $\mathcal{R}_3'$ decreases the objective by a concrete amount.

Proposition 14.7 *Suppose our current iterate $\boldsymbol{z}^{(r)} \in \mathcal{R}_3'$, and the trust-region subproblem takes a constrained step, i.e., the optimizer to (14.1.1) satisfies $\|\boldsymbol{\delta}_\star\| = \Delta$. We have the $\boldsymbol{\delta}_\star$ leads to*

$$f(\boldsymbol{z}^{(r+1)}) - f(\boldsymbol{z}^{(r)}) \le -d_4 \doteq -\frac{m_H^2 \Delta^2}{4 M_H}. \tag{14.2.11}$$

provided that

$$\Delta \le m_H^2 / (4 M_H L_h). \tag{14.2.12}$$

Here m_H and M_H are as defined in Lemma 14.2.

Proof See Section 18.8 on Page 185. ■

The next proposition shows that when Δ is properly tuned, an unconstrained step in $\mathcal{R}_3'$ dramatically reduces the norm of the gradient.

Proposition 14.8 (Quadratic Convergence of the Norm of the Gradient) *Suppose our current iterate $\boldsymbol{z}^{(r)} \in \mathcal{R}_3'$, and the trust-region subproblem takes an unconstrained step, i.e., the unique optimizer to (14.1.1) satisfies $\|\boldsymbol{\delta}_\star\| < \Delta$. We have the $\boldsymbol{\delta}_\star$ leads to $\boldsymbol{z}^{(r+1)} = \boldsymbol{z}^{(r)} + \boldsymbol{\delta}_\star$ that obeys*

$$\|\nabla f(\boldsymbol{z}^{(r+1)})\| \le \frac{1}{m_H^2}(L_h + \frac{32}{\|\boldsymbol{x}\|}M_H)\|\nabla f(\boldsymbol{z}^{(r)})\|^2, \tag{14.2.13}$$

provided

$$\Delta \le \|\boldsymbol{x}\|/10. \tag{14.2.14}$$

Here M_H and m_H are as defined in Lemma 14.2.

Proof See Section 18.9 on Page 186. ■

The next proposition shows that when Δ is properly tuned, as soon as an unconstrained $\mathcal{R}_3'$ step is taken, all future iterations take unconstrained $\mathcal{R}_3'$ steps. Moreover, the sequence converges quadratically to the target set X.

Proposition 14.9 (Quadratic Convergence of the Iterates in $\mathcal{R}_3'$) *Suppose the trust-region algorithm starts to take an unconstrained step in $\mathcal{R}_3'$ at $\boldsymbol{z}^{(r)}$ for a certain $r \in \mathbb{N}$. Then all future steps will be unconstrained steps in $\mathcal{R}_3'$, and*

$$\left\|\boldsymbol{h}(\boldsymbol{z}^{(r+r')})\right\| \le \frac{4\sqrt{2}m_H^2}{\|\boldsymbol{x}\|^2}\left(L_h + \frac{32}{\|\boldsymbol{x}\|}M_H\right)^{-1} 2^{-2^{r'}} \tag{14.2.15}$$

for all integers $r' \ge 1$, provided that

$$\Delta \le \min\left\{\frac{\|\boldsymbol{x}\|}{10}, \frac{m_H\|\boldsymbol{x}\|^2}{M_H\sqrt{40\sqrt{2}L_h(L_h + 32M_H/\|\boldsymbol{x}\|)}}, \frac{m_H^3}{\sqrt{2}M_H^2(L_h + 32M_H/\|\boldsymbol{x}\|)}\right\}. \tag{14.2.16}$$

Proof See Section 18.10 on Page 189. ■

Now we are ready to piece together the above technical propositions to prove our main algorithmic theorem.

Theorem 14.10 (TRM Convergence) *Suppose $m \ge Cn\log^3 n$ for a sufficiently large constant C. Then with probability at least $1 - c_a m^{-1}$, the trust-region algorithm with an* arbitrary initialization *$\boldsymbol{z}^{(0)} \in \mathbb{CB}^n(R_0)$,* *with $R_0 = 3(\frac{1}{m}\sum_{k=1}^m y_k^2)^{1/2}$, will return a solution that is ε-close to the target set X in*

$$\frac{c_b}{\Delta^2 \|\boldsymbol{x}\|^2} f(\boldsymbol{z}^{(0)}) + \log\log(\frac{c_c \|\boldsymbol{x}\|}{\varepsilon}) \tag{14.2.17}$$

steps, provided that

$$\Delta \le c_d (n^{7/2} \log^{7/2} m)^{-1} \|\boldsymbol{x}\| . \tag{14.2.18}$$

Here c_a through c_d are positive absolute constants.

Proof When $m \ge C_1 n \log^3 n$ for a sufficiently large constant C_1, the assumption of Theorem 13.2 is satisfied. Moreover, with probability at least $1 - c_2 m^{-1}$, the following estimates hold:

$$L_f = C_3 n^{5/2} \log^{5/2} m \|\boldsymbol{x}\|^3 , \quad L_g = C_3 n^2 \log^2 m \|\boldsymbol{x}\|^2 , \quad L_h = C_3 n^{3/2} \log^{3/2} m \|\boldsymbol{x}\| ,$$
$$m_H = 22/25 \|\boldsymbol{x}\|^2 , \quad M_H = 9/2 \|\boldsymbol{x}\|^2$$

for a certain positive absolute constant C_3. From the technical lemmas and propositions in Section 14.2.2, it can be verified that when

$$\Delta \le c_4 (n^{7/2} \log^{7/2} m)^{-1} \|\boldsymbol{x}\| ,$$

for a positive absolute constant c_4, all requirements on Δ are satisfied.

Write $\mathcal{R}_A \doteq \Gamma \setminus \mathcal{R}_3'$, where $\Gamma \doteq \mathbb{CB}^n(R_1)$ with $R_1 = 3\sqrt{n \log m} R_0$. Then a step in Γ is either a $\mathcal{R}_A$ or constrained $\mathcal{R}_3'$ step that reduces the objective value by a concrete amount, or an unconstrained $\mathcal{R}_3'$ step with all subsequent steps being unconstrained $\mathcal{R}_3'$. From discussion in Section 14.2.1, for an arbitrary initialization $\boldsymbol{z}^{(0)} \in \Gamma$, our choice of R_1 ensures that w.h.p. the sublevel set $\Pi \doteq \{\boldsymbol{z} : f(\boldsymbol{z}) \le f(\boldsymbol{z}^{(0)})\}$ is contained in Γ. Moreover, $\mathcal{R}_3'$ is also contained in Γ. $\mathcal{R}_A$ and constrained $\mathcal{R}_3'$ steps reduce the objective function, and therefore cannot cause the iterate sequence to leave Π. So $\mathcal{R}_A$ and constrained $\mathcal{R}_3'$ steps stay within Γ. Since unconstrained $\mathcal{R}_3'$ steps stay within $\mathcal{R}_3'$, they also stay within Γ, and the iterate sequence as a whole does not leave Γ.

In fact, the previous argument implies a generic iterate sequence consists of two phases: the first phase that takes consecutive $\mathcal{R}_A$ or constrained $\mathcal{R}_3'$ steps, and the second phase that takes consecutive unconstrained $\mathcal{R}_3'$ steps. Either of the two can be absent depending on the initialization and parameter setting for the TRM algorithm.

By Proposition 14.4, 14.5, 14.6, and 14.7, from $\boldsymbol{z}^{(0)}$ it takes at most

$$f(\boldsymbol{z}^{(0)})/\min(d_1, d_2, d_3, d_4)$$

steps for the iterate sequence to start take consecutive unconstrained $\mathcal{R}_3'$ step, or to stops on the target set X. In the former case, by Proposition 14.9, the sequence then takes at most

$$\log\log\left(\frac{4\sqrt{2}m_H^2}{(L_h + 32M_H/\left\|\boldsymbol{x}\right\|)\left\|\boldsymbol{x}\right\|^2\varepsilon}\right)$$

more steps to reach an ε-close point to the target set X.

In sum, the number of iterations to obtain an ε-close solution to the target set X can be grossly bounded by

$$\#\text{Iter} \ \le \ \frac{f(\boldsymbol{z}^{(0)})}{\min\left\{d_1, d_2, d_3, d_4\right\}} + \log\log\left(\frac{4\sqrt{2}m_H^2}{(L_h + 32M_H/\left\|\boldsymbol{x}\right\|)\left\|\boldsymbol{x}\right\|^2\varepsilon}\right).$$

Using our previous estimates of m_H, M_H, and L_H, and taking $\min\{d_1, d_2, d_3, d_4\} = c_5\Delta^2\left\|\boldsymbol{x}\right\|^2$, we arrive at the claimed result. ■

Chapter 15

Numerical Simulations

A theory can be proved by experiment; but no path leads from experiment to the birth of a theory.

Albert Einstein

In this chapter, we investigate experimentally the number of measurements m required to ensure that $f(\boldsymbol{z})$ is well-structured, in the sense of our theorems. This entails solving large instances of $f(\boldsymbol{z})$. To this end, we deploy the modified Manopt package.

We fix $n = 1,000$ and vary the ratio m/n from 4 to 10. For each m, we generate a fixed instance: a fixed signal $\boldsymbol{x}$, and a fixed set of complex Gaussian vectors. We run the TRM algorithm 25 times for each problem

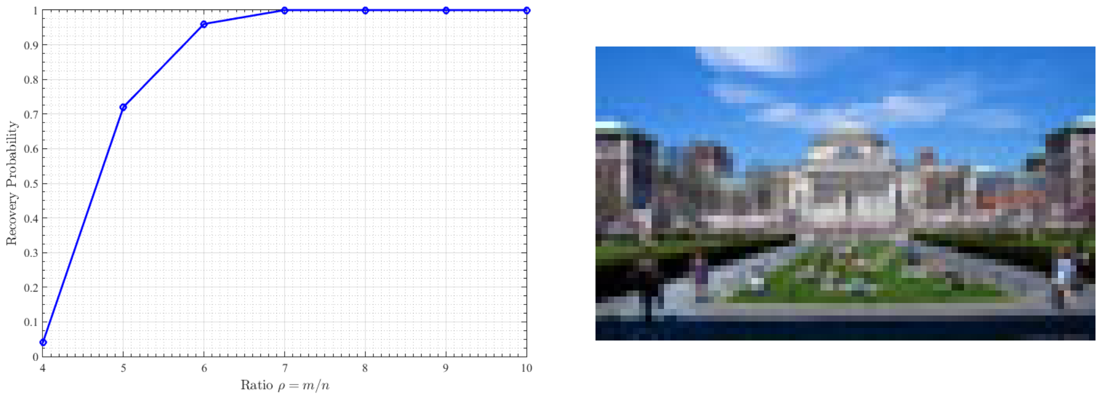


Figure 15.1: (Left) Recovery performance for GPR when optimizing (12.1.1) with the TRM. With $n = 1000$ and m varying, we consider a fixed problem instance for each m, and run the TRM algorithm 25 times from independently random initializations. The empirical recovery probability is a test of whether the benign geometric structure holds. (Right) A small "artistic" Columbia University campus image we use for comparing TRM and gradient descent.

instance, with independent random initializations. Successfully recovery is declared if at termination the optimization variable $\boldsymbol{z}_\infty$ satisfies

$$\varepsilon_{\mathrm{Rel}} \doteq \left\| \boldsymbol{z}_\infty - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z}_\infty)} \right\| / \left\| \boldsymbol{x} \right\| \leq 10^{-3}.$$

The recovery probability is empirically estimated from the 25 repetitions for each m. Intuitively, when the recovery probability is below one, there are spurious local minimizers. In this case, the number of samples m is not large enough to ensure the finite-sample function landscape $f(\boldsymbol{z})$ to be qualitatively the same as the asymptotic version $\mathbb{E}_{\boldsymbol{a}}[f(\boldsymbol{z})]$. Figure 15.1 shows the recovery performance. It seems that $m = 7n$ samples may be sufficient to ensure the geometric property holds.[1] On the other hand, $m = 6n$ is not sufficient, whereas in theory it is known $4n$ samples are enough to guarantee measurement injectivity for complex signals [BCE06].[2]

We now briefly compare TRM and gradient descent in terms of running time. We take a small ($n = 80 \times 47$) image of Columbia University campus (Figure 15.1 (Right)), and make $m = 5n \log n$ complex Gaussian measurements. The TRM solver is the same as above, and the gradient descent solver is one with backtracking line search. We repeat the experiment 10 times, with independently generated random measurements and initializations each time. On average, the TRM solver returns a solution with $\varepsilon_{\mathrm{Rel}} \leq 10^{-4}$ in about 2600 seconds, while the gradient descent solver produces a solution with $\varepsilon_{\mathrm{Rel}} \sim 10^{-2}$ in about 6400 seconds. The point here is not to exhaustively benchmark the two – they both involve many implementation details and tuning parameters and they have very different memory requirements. It is just to suggest that second-order methods can be implemented in a practical manner for large-scale GPR problems.[3]

[1] This prescription should be taken with a grain of salt, as here we have only tested a single fixed n.

[2] Numerics in [CC15] suggest that under the same measurement model, $m = 5n$ is sufficient for efficient recovery. Our requirement on control of the whole function landscape and hence “initialization-free" algorithm may need the additional complexity.

[3] The main limitation in this experiment was not the TRM solver, but the need to store the vectors $\boldsymbol{a}_1, \ldots \boldsymbol{a}_m$. For other measurement models, such as the coded diffraction model [CLS15a], “matrix-free” calculation is possible, and storage is no longer a bottleneck.

Chapter 16

Discussion

> The best way to predict the future is to invent it.
>
> Allan Kay

In this work, we provide a complete geometric characterization of the nonconvex formulation (12.1.1) for the GPR problem. The benign geometric structure allows us to design a second-order trust-region algorithm that efficiently finds a global minimizer of (12.1.1), without special initialization. We close the main body of this work by discussing possible extensions and relevant open problems.

Sample complexity and measurement schemes. Our result (Theorem 13.2 and Theorem 14.10) indicates that $m \geq C_1 n \log^3 n$ samples are sufficient to guarantee the favorable geometric property and efficient recovery, while our simulations suggested that $C_2 n \log n$ or even $C_3 n$ is enough. For efficient recovery only, $m \geq C_4 n$ are known to be sufficient [CC15] (and uniformly for all signals; see also [CLS15b]). It is interesting to see if the gaps can be closed. Our current analysis pertains to i.i.d. Gaussian measurements only, which are not practical. It is important to extend the geometric analysis to more practical measurement schemes, such as t-designs [GKK13] and masked Fourier transform measurements [CLS15a]. A preliminary study of the low-dimensional function landscape for the latter scheme produces very positive result; see Figure 16.1.

Sparse phase retrieval. A special case of GPR is when the underlying signal $\boldsymbol{x}$ is known to be sparse, which can be considered as a quadratic compressed sensing problem [OYVS13, OYDS13, OYDS12, LV13, JOH13, SBE14]. Since $\boldsymbol{x}$ is sparse, the lifted matrix $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^*$ is sparse and has rank one. Thus, existing convex relaxation methods [OYVS13, OYDS13, LV13, JOH13] formulated it as a simultaneously low-rank and sparse

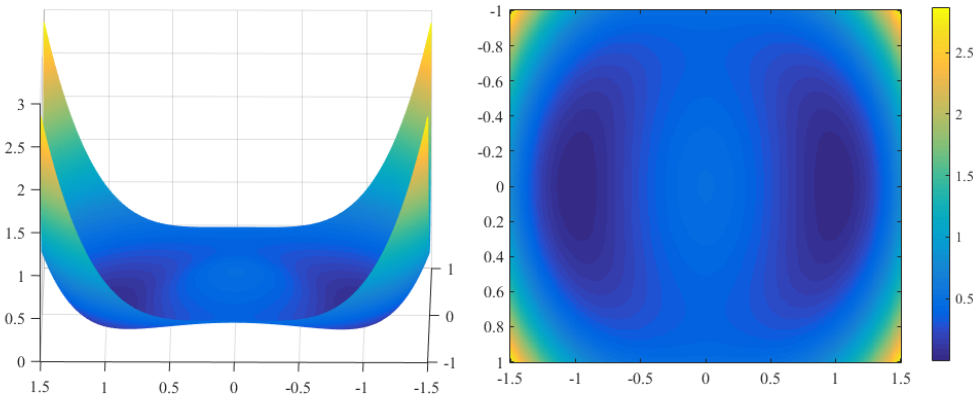


Figure 16.1: Function landscape of (12.1.1) for $\boldsymbol{x} = [1; 0]$ and $m \to \infty$ for the masked Fourier transform measurements (coded diffraction model [CLS15a]). The landscape is qualitatively similar to that for the Gaussian model (Figure 12.2).

recovery problem. For the latter problem, however, known convex relaxations are suboptimal [OJF$^+$12, MHWG14]. Let k be the number of nonzeros in the target signal. [LV13, JOH13] showed that natural convex relaxations require $C_5 k^2 \log n$ samples for correct recovery, instead of the optimal order $O(k \log(n/k))$. A similar gap is also observed with certain nonconvex methods [CLM15]. It is tempting to ask whether novel nonconvex formulations and analogous geometric analysis as taken here could shed light on this problem.

Other structured nonconvex problems. We have mentioned recent surge of works on provable nonconvex heuristics [JNS13, Har14, HW14, NNS$^+$14, JN14, SL14, JO14, WCCL15, SRO15, ZL15, TBSR15, CW15, AGJ14a, AGJ14b, AJSN15, GHJY15, QSW14, HSSS15, AAJ$^+$13, AGM13, AAN13, ABGM14, AGMM15, SQW15a, YCS13, SA14c, LWB13, LJ15, LLJB15, EW15, Bou16, JJKN15]. While the initialization plus local refinement analyses generally produce interesting theoretical results, they do not explain certain empirical successes that do not rely on special initializations. The geometric structure and analysis we work with in our recent work [SQW15a, SQW15b] (see also [GHJY15] and [AG16]) seem promising in this regard. It is interesting to consider whether analogous geometric structure exists for other practical problems.

Chapter 17

Proofs of Technical Results for Function Landscape

The idea of concentration of measure (which was discovered by V. Milman) is arguably one of the great ideas of analysis in our times. While its impact on Probability is only a small part of the whole picture, this impact should not be ignored.

Micheal Talagrand, in *A new look at independence*

17.1 Auxiliary lemmas

Lemma 17.1 *For the function* $f(\boldsymbol{z}) : \mathbb{C}^n \mapsto \mathbb{R}$ *defined in* (12.1.1)*, we have*

$$\mathbb{E}\left[f(\boldsymbol{z})\right] = \left\|\boldsymbol{x}\right\|^4 + \left\|\boldsymbol{z}\right\|^4 - \left\|\boldsymbol{x}\right\|^2 \left\|\boldsymbol{z}\right\|^2 - \left|\boldsymbol{x}^* \boldsymbol{z}\right|^2, \tag{17.1.1}$$

$$\nabla \mathbb{E}\left[f(\boldsymbol{z})\right] = \begin{bmatrix} \nabla_{\boldsymbol{z}} \mathbb{E}\left[f(\boldsymbol{z})\right] \\ \nabla_{\overline{\boldsymbol{z}}} \mathbb{E}\left[f(\boldsymbol{z})\right] \end{bmatrix} = \begin{bmatrix} \left(2\left\|\boldsymbol{z}\right\|^2 \boldsymbol{I} - \left\|\boldsymbol{x}\right\|^2 \boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^*\right) \boldsymbol{z} \\ \left(2\left\|\boldsymbol{z}\right\|^2 \boldsymbol{I} - \left\|\boldsymbol{x}\right\|^2 \boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^*\right) \overline{\boldsymbol{z}} \end{bmatrix}, \tag{17.1.2}$$

$$\nabla^2 \mathbb{E}\left[f(\boldsymbol{z})\right] = \begin{bmatrix} 2\boldsymbol{z}\boldsymbol{z}^* - \boldsymbol{x}\boldsymbol{x}^* + \left(2\left\|\boldsymbol{z}\right\|^2 - \left\|\boldsymbol{x}\right\|^2\right)\boldsymbol{I} & 2\boldsymbol{z}\boldsymbol{z}^\top \\ 2\overline{\boldsymbol{z}}\boldsymbol{z}^* & 2\overline{\boldsymbol{z}}\boldsymbol{z}^\top - \overline{\boldsymbol{x}}\boldsymbol{x}^\top + \left(2\left\|\boldsymbol{z}\right\|^2 - \left\|\boldsymbol{x}\right\|^2\right)\boldsymbol{I} \end{bmatrix}. \tag{17.1.3}$$

Proof By definition (12.1.1), notice that

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{z})\right] &= \frac{1}{2}\mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}(n)}\left[\left(\left|\langle \boldsymbol{a},\boldsymbol{x}\rangle\right|^2 - \left|\langle \boldsymbol{a},\boldsymbol{z}\rangle\right|^2\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}(n)}\left[\left|\langle \boldsymbol{a},\boldsymbol{x}\rangle\right|^4\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}(n)}\left[\left|\langle \boldsymbol{a},\boldsymbol{z}\rangle\right|^4\right] - \mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}(n)}\left[\left|\langle \boldsymbol{a},\boldsymbol{x}\rangle\right|^2\left|\langle \boldsymbol{a},\boldsymbol{z}\rangle\right|^2\right].\end{aligned}$$

We now evaluate the three terms separately. Note that the law $\mathcal{CN}(n)$ is invariant to unitary transform. Thus,

$$\mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}(n)}\left[\left|\langle \boldsymbol{a},\boldsymbol{x}\rangle\right|^4\right] = \mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}(n)}\left[\left|\langle \boldsymbol{a},\boldsymbol{e}_1\rangle\right|^4\right]\left\|\boldsymbol{x}\right\|^4 = \mathbb{E}_{a\sim\mathcal{N}(0,1/2)+\mathrm{i}\,\mathcal{N}(0,1/2)}\left[|a|^4\right]\left\|\boldsymbol{x}\right\|^4 = 2\left\|\boldsymbol{x}\right\|^4.$$

Similarly, we also obtain $\mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}}\left[\left|\langle \boldsymbol{a},\boldsymbol{z}\rangle\right|^4\right] = 2\left\|\boldsymbol{z}\right\|^4$. Now for the cross term,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}(n)}\left[\left|\langle \boldsymbol{a},\boldsymbol{x}\rangle\right|^2\left|\langle \boldsymbol{a},\boldsymbol{z}\rangle\right|^2\right] \\ &= \mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}(n)}\left[\left|\langle \boldsymbol{a},\boldsymbol{e}_1\rangle\right|^2\left|\left\langle \boldsymbol{a}, s_1\mathrm{e}^{\mathrm{i}\phi_1}\boldsymbol{e}_1 + s_2\mathrm{e}^{\mathrm{i}\phi_2}\boldsymbol{e}_2\right\rangle\right|^2\right]\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 \quad [\text{where } s_1^2+s_2^2=1] \\ &= \mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}(n)}\left[\left|a_1\right|^2\left|s_1\overline{a_1}\mathrm{e}^{\mathrm{i}\phi_1} + s_2\overline{a_2}\mathrm{e}^{\mathrm{i}\phi_2}\right|^2\right]\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 \\ &= \mathbb{E}_{\boldsymbol{a}\sim\mathcal{CN}(n)}\left[\left|a_1\right|^2\left(s_1^2\left|a_1\right|^2 + s_2^2\left|a_2\right|^2\right)\right]\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 \\ &= \left(1+s_1^2\right)\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 = \left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 + \left|\boldsymbol{x}^*\boldsymbol{z}\right|^2.\end{aligned}$$

Gathering the above results, we obtain (17.1.1). By taking Wirtinger derivative (12.5.2) with respect to (17.1.1), we obtain the Wirtinger gradient and Hessian in (17.1.2), (17.1.3) as desired. ■

> **Lemma 17.2** *For $\boldsymbol{a}\sim\mathcal{CN}(n)$ and any fixed vector $\boldsymbol{v}\in\mathbb{C}^n$, it holds that*
>
> $$\mathbb{E}\left[\left|\boldsymbol{a}^*\boldsymbol{v}\right|^2\boldsymbol{a}\boldsymbol{a}^*\right] = \boldsymbol{v}\boldsymbol{v}^* + \left\|\boldsymbol{v}\right\|^2\boldsymbol{I}, \quad \textit{and} \quad \mathbb{E}\left[\left(\boldsymbol{a}^*\boldsymbol{v}\right)^2\boldsymbol{a}\boldsymbol{a}^\top\right] = 2\boldsymbol{v}\boldsymbol{v}^\top.$$

Proof Observe that for $i\neq j$,

$$\boldsymbol{e}_i^*\mathbb{E}\left[\left|\boldsymbol{a}^*\boldsymbol{v}\right|^2\boldsymbol{a}\boldsymbol{a}^*\right]\boldsymbol{e}_j = \sum_{q,\ell}\mathbb{E}\left[\overline{a(q)}a(\ell)v(q)\overline{v(\ell)}a(i)\overline{a(j)}\right] = \mathbb{E}\left[\left|a(i)\right|^2\left|a(j)\right|^2\right]v(i)\overline{v(j)} = v(i)\overline{v(j)}.$$

Similarly,

$$\begin{aligned}\boldsymbol{e}_i^*\mathbb{E}\left[\left|\boldsymbol{a}^*\boldsymbol{v}\right|^2\boldsymbol{a}\boldsymbol{a}^*\right]\boldsymbol{e}_i &= \sum_{q,\ell}\mathbb{E}\left[\overline{a(q)}a(\ell)v(q)\overline{v(\ell)}\left|a(i)\right|^2\right] \\ &= \mathbb{E}\left[\left|a(i)\right|^4\left|v(i)\right|^2\right] + \sum_{q\neq i}\mathbb{E}\left[\left|a(q)\right|^2\left|v(q)\right|^2\left|a(i)\right|^2\right] = \left|v(i)\right|^2 + \left\|\boldsymbol{v}\right\|^2.\end{aligned}$$

Similar calculation yields the second expectation. ■

Lemma 17.3 *Let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_m$ be i.i.d. copies of $\boldsymbol{a} \sim \mathcal{CN}(n)$. For any $\delta \in (0,1)$ and any $\boldsymbol{v} \in \mathbb{C}^n$, when $m \geq C(\delta) n \log n$, we have that with probability at least $1 - c_a \delta^{-2} m^{-1} - c_b \exp\left(-c_c \delta^2 m / \log m\right)$*

$$\left\| \frac{1}{m} \sum_{k=1}^m \left|\boldsymbol{a}_k^* \boldsymbol{v}\right|^2 \boldsymbol{a}_k \boldsymbol{a}_k^* - \left(\boldsymbol{v}\boldsymbol{v}^* + \left\|\boldsymbol{v}\right\|^2 \boldsymbol{I}\right) \right\| \leq \delta \left\|\boldsymbol{v}\right\|^2,$$

$$\left\| \frac{1}{m} \sum_{k=1}^m \left(\boldsymbol{a}_k^* \boldsymbol{v}\right)^2 \boldsymbol{a}_k \boldsymbol{a}_k^\top - 2\boldsymbol{v}\boldsymbol{v}^\top \right\| \leq \delta \left\|\boldsymbol{v}\right\|^2.$$

Here $C(\delta)$ is a constant depending on δ and c_a, c_b and c_c are positive absolute constants.

Proof We work out the results on $\frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^* \boldsymbol{v}\right|^2 \boldsymbol{a}_k \boldsymbol{a}_k^*$ first. By the unitary invariance of the Gaussian measure and rescaling, it is enough to consider $\boldsymbol{v} = \boldsymbol{e}_1$. We partition each vector $\boldsymbol{a}_k$ as $\boldsymbol{a}_k = [a_k(1); \widetilde{\boldsymbol{a}}_k]$ and upper bound the target quantity as:

$$\begin{aligned} & \left\| \frac{1}{m} \sum_{k=1}^m \left|a_k(1)\right|^2 \begin{bmatrix} \left|a_k(1)\right|^2 & a_k(1)\widetilde{\boldsymbol{a}}_k^* \\ \overline{a_k(1)}\widetilde{\boldsymbol{a}}_k & \widetilde{\boldsymbol{a}}_k \widetilde{\boldsymbol{a}}_k^* \end{bmatrix} - \left(\boldsymbol{e}_1 \boldsymbol{e}_1^* + \boldsymbol{I}\right) \right\| \\ \leq & \left| \frac{1}{m} \sum_{k=1}^m \left(\left|a_k(1)\right|^4 - 2 \right) \right| + \left\| \frac{1}{m} \sum_{k=1}^m \left|a_k(1)\right|^2 \begin{bmatrix} 0 & a_k(1)\widetilde{\boldsymbol{a}}_k^* \\ \overline{a_k(1)}\widetilde{\boldsymbol{a}}_k & \boldsymbol{0} \end{bmatrix} \right\| \\ & + \left\| \frac{1}{m} \sum_{k=1}^m \left|a_k(1)\right|^2 \left(\widetilde{\boldsymbol{a}}_i \widetilde{\boldsymbol{a}}_k^* - \boldsymbol{I}_{n-1}\right) \right\| + \left| \frac{1}{m} \sum_{k=1}^m \left(\left|a_k(1)\right|^2 - 1 \right) \right|. \end{aligned}$$

By Chebyshev's inequality, we have with probability at least $1 - c_1 \delta^{-2} m^{-1}$,

$$\left| \frac{1}{m} \sum_{k=1}^m \left(\left|a_k(1)\right|^4 - 2 \right) \right| \leq \frac{\delta}{4} \quad \text{and} \quad \left| \frac{1}{m} \sum_{k=1}^m \left(\left|a_k(1)\right|^2 - 1 \right) \right| \leq \frac{\delta}{4}.$$

To bound the second term, we note that

$$\begin{aligned} \left\| \frac{1}{m} \sum_{k=1}^m \left|a_k(1)\right|^2 \begin{bmatrix} 0 & a_k(1)\widetilde{\boldsymbol{a}}_k^* \\ \overline{a_k(1)}\widetilde{\boldsymbol{a}}_k & \boldsymbol{0} \end{bmatrix} \right\| & = \left\| \frac{1}{m} \sum_{k=1}^m \left|a_k(1)\right|^2 a_k(1)\widetilde{\boldsymbol{a}}_k^* \right\| \\ & = \sup_{\boldsymbol{w} \in \mathbb{C}^{n-1}: \left\|\boldsymbol{w}\right\| = 1} \frac{1}{m} \sum_{k=1}^m \left|a_k(1)\right|^2 a_k(1)\widetilde{\boldsymbol{a}}_k^* \boldsymbol{w}. \end{aligned}$$

For all $\boldsymbol{w}$ and all $k \in [m]$, $\widetilde{\boldsymbol{a}}_k^* \boldsymbol{w}$ is distributed as $\mathcal{CN}(1)$ that is independent of the $\{a_k(1)\}$ sequence. So for one realization of $\{a_k(1)\}$, the Hoeffding-type inequality of Lemma A.4 implies

$$\mathbb{P}\left[\frac{1}{m} \sum_{k=1}^m \left|a_k(1)\right|^2 a_k(1)\widetilde{\boldsymbol{a}}_k^* \boldsymbol{w} > t \right] \leq \mathrm{e}\exp\left(-\frac{c_2 m^2 t^2}{\sum_{k=1}^m \left|a_k(1)\right|^6} \right),$$

for any $\boldsymbol{w}$ with $\left\|\boldsymbol{w}\right\| = 1$ and any $t > 0$. Taking $t = \delta/8$, together with a union bound on a $1/2$-net on the sphere, we obtain

$$\mathbb{P}\left[\left\|\frac{1}{m}\sum_{k=1}^{m}\left|a_k(1)\right|^2 a_k(1)\widetilde{\boldsymbol{a}}_k^*\right\| > \delta/4\right] \le \mathrm{e}\exp\left(-\frac{c_2 m^2\delta^2}{64\sum_{k=1}^{m}\left|a_k(1)\right|^6} + 12(n-1)\right).$$

Now an application of Chebyshev's inequality gives that $\sum_{k=1}^{m}\left|a_k(1)\right|^6 \le 20m$ with probability at least $1 - c_3 m^{-1}$. Substituting this into the above, we conclude that whenever $m \ge C_4\delta^{-2}n$ for some sufficiently large C_4,

$$\left\|\frac{1}{m}\sum_{k=1}^{m}\left|a_k(1)\right|^2 a_k(1)\widetilde{\boldsymbol{a}}_k^*\right\| \le \delta/4$$

with probability at least $1 - c_3 m^{-1} - \exp\left(-c_5\delta^2 m\right)$.

To bound the third term, we note that

$$\left\|\frac{1}{m}\sum_{k=1}^{m}\left|a_k(1)\right|^2\left(\widetilde{\boldsymbol{a}}_k\widetilde{\boldsymbol{a}}_k^* - \boldsymbol{I}_{n-1}\right)\right\| = \sup_{\boldsymbol{w}\in\mathbb{C}^{n-1}:\|\boldsymbol{w}\|=1}\frac{1}{m}\sum_{k=1}^{m}\left|a_k(1)\right|^2\left(\left|\widetilde{\boldsymbol{a}}_k^*\boldsymbol{w}\right|^2 - 1\right).$$

For all fixed $\boldsymbol{w}$ and all $k \in [m]$, $\widetilde{\boldsymbol{a}}_k^*\boldsymbol{w} \sim \mathcal{CN}(1)$. Thus, $\left|\widetilde{\boldsymbol{a}}_k^*\boldsymbol{w}\right|^2 - 1$ is centered sub-exponential. So for one realization of $\{a_k(1)\}$, Bernstein's inequality (Lemma A.5) implies

$$\mathbb{P}\left[\frac{1}{m}\sum_{k=1}^{m}\left|a_k(1)\right|^2\left(\left|\widetilde{\boldsymbol{a}}_k^*\boldsymbol{w}\right|^2 - 1\right) > t\right] \le 2\exp\left(-c_6\min\left(\frac{t^2}{c_7^2\sum_{k=1}^{m}\left|a_k(1)\right|^4}, \frac{t}{c_7\max_{i\in[m]}\left|a_k(1)\right|^2}\right)\right)$$

for any fixed $\boldsymbol{w}$ with $\|\boldsymbol{w}\| = 1$ and any $t > 0$. Taking $t = \delta/8$, together with a union bound on a $1/2$-net on the sphere, we obtain

$$\mathbb{P}\left[\left\|\frac{1}{m}\sum_{k=1}^{m}\left|a_k(1)\right|^2\left(\widetilde{\boldsymbol{a}}_k\widetilde{\boldsymbol{a}_k}^* - \boldsymbol{I}_{n-1}\right)\right\| > \frac{\delta}{4}\right]$$
$$\le 2\exp\left(-c_6\min\left(\frac{m^2\delta^2/64}{c_7^2\sum_{k=1}^{m}\left|a_k(1)\right|^4}, \frac{m\delta/8}{c_7\max_{i\in[m]}\left|a_k(1)\right|^2}\right) + 12(n-1)\right).$$

Chebyshev's inequality and the union bound give that

$$\sum_{k=1}^{m}\left|a_k(1)\right|^4 \le 10m, \quad \text{and} \quad \max_{i\in[m]}\left|a_k(1)\right|^2 \le 10\log m$$

hold with probability at least $1 - c_8 m^{-1} - m^{-4}$. To conclude, when $m \ge C_9(\delta)\delta^{-2}n\log n$ for some sufficiently large constant $C_9(\delta)$,

$$\left\|\frac{1}{m}\sum_{k=1}^{m}\left|a_k(1)\right|^2\left(\widetilde{\boldsymbol{a}}_k\widetilde{\boldsymbol{a}_k}^* - \boldsymbol{I}_{n-1}\right)\right\| \le \frac{\delta}{4}$$

with probability at least $1 - c_8 m^{-1} - m^{-4} - 2\exp\left(-c_{10}\delta^2 m/\log m\right)$.

Collecting the above bounds and probabilities yields the claimed results. Similar arguments prove the claim on $\frac{1}{m}\sum_{k=1}^m \left(\boldsymbol{a}_k^* \boldsymbol{v}\right) \boldsymbol{a}_k \boldsymbol{a}_k^\top$ also, completing the proof. ■

> **Lemma 17.4** *Let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_m$ be i.i.d. copies of $\boldsymbol{a} \sim \mathcal{CN}(n)$. For any $\delta \in (0,1)$, when $m \ge C(\delta) n \log n$, it holds with probability at least $1 - c' \exp\left(-c(\delta) m\right) - c'' m^{-n}$ that*
>
> $$\frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 \left|\boldsymbol{a}_k^* \boldsymbol{w}\right|^2 \ge (1-\delta)\left(\left\|\boldsymbol{w}\right\|^2 \left\|\boldsymbol{z}\right\|^2 + \left|\boldsymbol{w}^* \boldsymbol{z}\right|^2\right) \quad \textit{for all } \boldsymbol{z}, \boldsymbol{w} \in \mathbb{C}^n,$$
>
> $$\frac{1}{m}\sum_{k=1}^m \left[\Re(\boldsymbol{a}_k^* \boldsymbol{z})(\boldsymbol{w}^* \boldsymbol{a}_k)\right]^2 \ge (1-\delta)\left(\frac{1}{2}\left\|\boldsymbol{z}\right\|^2 \left\|\boldsymbol{w}\right\|^2 + \frac{3}{2}[\Re \boldsymbol{z}^* \boldsymbol{w}]^2 - \frac{1}{2}[\Im \boldsymbol{z}^* \boldsymbol{w}]^2\right) \quad \textit{for all } \boldsymbol{z}, \boldsymbol{w} \in \mathbb{C}^n.$$
>
> *Here $C(\delta)$ and $c(\delta)$ are constants depending on δ and c' and c'' are positive absolute constants.*

Proof By Lemma 17.2, $\mathbb{E}\left[\left|\boldsymbol{a}^* \boldsymbol{w}\right|^2 \left|\boldsymbol{a}^* \boldsymbol{z}\right|^2\right] = \left\|\boldsymbol{w}\right\|^2 \left\|\boldsymbol{z}\right\|^2 + \left|\boldsymbol{w}^* \boldsymbol{z}\right|^2$. By homogeneity, it is enough to prove the result for all $\boldsymbol{w}, \boldsymbol{z} \in \mathbb{CS}^{n-1}$. For a pair of fixed $\boldsymbol{w}, \boldsymbol{z} \in \mathbb{CS}^{n-1}$, Lemma A.6 implies that for any $\delta \in (0,1)$,

$$\sum_{k=1}^m \left|\boldsymbol{a}_k^* \boldsymbol{w}\right|^2 \left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 \ge \left(1 - \frac{\delta}{2}\right) m \left(1 + \left|\boldsymbol{w}^* \boldsymbol{z}\right|^2\right)$$

with probability at least $1 - \exp(-c_1 \delta^2 m)$. For a certain $\varepsilon \in (0,1)$ to be fixed later and an ε-net $N_\varepsilon^1 \times N_\varepsilon^2$ that covers $\mathbb{CS}^{n-1} \times \mathbb{CS}^{n-1}$, we have that the event

$$\mathcal{E}_0 \doteq \left\{\sum_{k=1}^m \left|\boldsymbol{a}_k^* \boldsymbol{w}\right|^2 \left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 \ge \left(1 - \frac{\delta}{2}\right) m \left(1 + \left|\boldsymbol{w}^* \boldsymbol{z}\right|^2\right) \quad \forall\, \boldsymbol{w}, \boldsymbol{z} \in N_\varepsilon^1 \times N_\varepsilon^2\right\}$$

holds with probability at least $1 - \exp\left(-c_1 \delta^2 m + 4n \log(3/\varepsilon)\right)$ by a simple union bound. Now conditioned on $\mathcal{E}_0$, we have for every $\boldsymbol{z} \in \mathbb{CS}^{n-1}$ can be written as $\boldsymbol{z} = \boldsymbol{z}_0 + \boldsymbol{e}$ for certain $\boldsymbol{z}_0 \in N_\varepsilon^1$ and $\boldsymbol{e}$ with $\left\|\boldsymbol{e}\right\| \le \varepsilon$; similarly $\boldsymbol{w} = \boldsymbol{w}_0 + \boldsymbol{\zeta}$ for $\boldsymbol{w}_0 \in N_\varepsilon^2$ and $\left\|\boldsymbol{\zeta}\right\| \le \varepsilon$. For the function $g(\boldsymbol{w}, \boldsymbol{z}) \doteq \sum_{k=1}^m \left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 \left|\boldsymbol{a}_k^* \boldsymbol{w}\right|^2$, with high probability,

$$\left\|\frac{\partial g}{\partial \boldsymbol{w}}\right\| = \left\|\sum_{k=1}^m \left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 \boldsymbol{w}^* \boldsymbol{a}_k \overline{\boldsymbol{a}}_k\right\| \le \left\|\boldsymbol{z}\right\|^2 \left\|\boldsymbol{w}\right\| \left\|\sum_{k=1}^m \left\|\boldsymbol{a}_k\right\|^2 \boldsymbol{a}_k \boldsymbol{a}_k^*\right\| \le 10 mn \sqrt{\log m},$$

$$\left\|\frac{\partial g}{\partial \boldsymbol{z}}\right\| = \left\|\sum_{k=1}^m \left|\boldsymbol{a}_k^* \boldsymbol{w}\right|^2 \boldsymbol{z}^* \boldsymbol{a}_k \overline{\boldsymbol{a}}_k\right\| \le \left\|\boldsymbol{w}\right\|^2 \left\|\boldsymbol{z}\right\| \left\|\sum_{k=1}^m \left\|\boldsymbol{a}_k\right\|^2 \boldsymbol{a}_k \boldsymbol{a}_k^*\right\| \le 10 mn \sqrt{\log m},$$

as $\max_{k \in [m]} \left\|\boldsymbol{a}_k\right\|^2 \le 5n \log m$ with probability at least $1 - c_2 m^{-n}$, and $\left\|\sum_{k=1}^m \boldsymbol{a}_k \boldsymbol{a}_k^*\right\| \le 2m$ with probability at least $1 - \exp(-c_3 m)$. Thus,

$$\sum_{k=1}^m \left|\boldsymbol{a}_k^* \boldsymbol{z}\right|^2 \left|\boldsymbol{a}_k^* \boldsymbol{w}\right|^2 \ge \left(1 - \frac{\delta}{3}\right) m - 40 \varepsilon mn \log m + \left(1 - \frac{\delta}{3}\right) m \left(\left|\boldsymbol{w}_0^* \boldsymbol{z}_0\right|^2 - 4\varepsilon\right).$$

Taking $\varepsilon = c_4(\delta)/(n \log m)$ for a sufficiently small $c_4(\delta) > 0$, we obtain that with probability at least $1 -$

$\exp\left(-c_1\delta^2 m + 4n\log(3n\log m/c_4(\delta))\right) - c_5 m^{-n}$,

$$\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 \left|\boldsymbol{a}_k^*\boldsymbol{w}\right|^2 \geq \left(1-\frac{2}{3}\delta\right) m \left(1+\left|\boldsymbol{w}_0^*\boldsymbol{z}_0\right|^2\right).$$

which, together with continuity of the function $(\boldsymbol{w},\boldsymbol{z}) \mapsto \left|\boldsymbol{w}^*\boldsymbol{z}\right|^2$, implies

$$\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 \left|\boldsymbol{a}_k^*\boldsymbol{w}\right|^2 \geq (1-\delta)\, m \left(1+\left|\boldsymbol{w}^*\boldsymbol{z}\right|^2\right).$$

It is enough to take $m \geq C_6\delta^{-2} n\log n$ to ensure the desired event happens with high probability.

To show the second inequality, first notice that $\mathbb{E}\left[\Re(\boldsymbol{a}_k^*\boldsymbol{z})(\boldsymbol{w}^*\boldsymbol{a}_k)\right]^2 = \frac{1}{2}\left\|\boldsymbol{z}\right\|^2\left\|\boldsymbol{w}\right\|^2 + \frac{3}{2}[\Re \boldsymbol{z}^*\boldsymbol{w}]^2 - \frac{1}{2}[\Im \boldsymbol{z}^*\boldsymbol{w}]^2$. The argument then proceeds to apply the discretization trick as above. ■

17.2 Proof of Proposition 13.3

Proof Direct calculation shows that

$$\begin{aligned}
&\begin{bmatrix} \boldsymbol{x}e^{i\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}}e^{-i\phi(\boldsymbol{z})} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix} \boldsymbol{x}e^{i\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}}e^{-i\phi(\boldsymbol{z})} \end{bmatrix} \\
&= \frac{1}{m}\sum_{k=1}^m \left(4\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2 - 2\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^4 + 2\Re\left[(\boldsymbol{a}_k^*\boldsymbol{z})^2(\boldsymbol{x}^*\boldsymbol{a}_k)^2 e^{-2i\phi(\boldsymbol{z})}\right]\right) \\
&= \frac{1}{m}\sum_{k=1}^m \left(2\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2 - 2\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^4\right) \\
&\quad + \frac{1}{m}\sum_{k=1}^m \left(2\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2 + 2\Re\left[(\boldsymbol{a}_k^*\boldsymbol{z})^2(\boldsymbol{x}^*\boldsymbol{a}_k)^2 e^{-2i\phi(\boldsymbol{z})}\right]\right).
\end{aligned}$$

Lemma 17.3 implies that when $m \geq C_1 n\log n$, with high probability,

$$\frac{2}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 \leq 2\left|\boldsymbol{x}^*\boldsymbol{z}\right|^2 + \frac{401}{200}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2.$$

On the other hand, by Lemma A.6, we have that

$$\frac{2}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^4 \geq \frac{399}{100}\left\|\boldsymbol{x}\right\|^4$$

holds with probability at least $1-\exp(-c_2 m)$. For the second summation, we have

$$\frac{1}{m}\sum_{k=1}^m \left(2\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2 + 2\Re\left[(\boldsymbol{a}_k^*\boldsymbol{z})^2(\boldsymbol{x}^*\boldsymbol{a}_k)^2 e^{-2i\phi(\boldsymbol{z})}\right]\right)$$

$$= \begin{bmatrix} \boldsymbol{z} \\ \overline{\boldsymbol{z}} \end{bmatrix}^* \nabla^2 f(\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})}) \begin{bmatrix} \boldsymbol{z} \\ \overline{\boldsymbol{z}} \end{bmatrix} \le \begin{bmatrix} \boldsymbol{z} \\ \overline{\boldsymbol{z}} \end{bmatrix}^* \nabla^2 \mathbb{E}\left[f(\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})})\right] \begin{bmatrix} \boldsymbol{z} \\ \overline{\boldsymbol{z}} \end{bmatrix} + \frac{1}{200}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 \le 6\left|\boldsymbol{x}^*\boldsymbol{z}\right|^2 + \frac{401}{200}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2,$$

with high probability, provided $m \ge C_3 n \log n$, according to Lemma 17.3.

Collecting the above estimates, we have that when $m \ge C_4 n \log n$ for a sufficiently large constant C_4, with high probability,

$$\begin{bmatrix} \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix} \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} \\ \overline{\boldsymbol{x}}\mathrm{e}^{-\mathrm{i}\phi(\boldsymbol{z})} \end{bmatrix} \le \frac{401}{100}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 + 8\left|\boldsymbol{x}^*\boldsymbol{z}\right|^2 - \frac{399}{100}\left\|\boldsymbol{x}\right\|^4 \le -\frac{1}{100}\left\|\boldsymbol{x}\right\|^4$$

for all $\boldsymbol{z} \in \mathcal{R}_1$. Dividing both sides of the above by $\left\|\boldsymbol{x}\right\|^2$ gives the claimed results. ■

17.3 Proof of Proposition 13.4

Proof Note that

$$\boldsymbol{z}^*\nabla_{\boldsymbol{z}} f(\boldsymbol{z}) = \frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^4 - \frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2.$$

By Lemma 17.4, when $m \ge C_1 n \log n$ for some sufficiently large C_1, with high probability,

$$\frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^4 \ge \frac{199}{100}\left\|\boldsymbol{z}\right\|^4$$

for all $\boldsymbol{z} \in \mathbb{C}^n$. On the other hand, Lemma 17.3 implies that when $m \ge C_2 n \log n$ for some sufficiently large C_2, with high probability,

$$\frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 \le \left|\boldsymbol{x}^*\boldsymbol{z}\right|^2 + \frac{1001}{1000}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2.$$

for all $\boldsymbol{z} \in \mathbb{C}^n$. Combining the above estimates, we have that when $m \ge \max(C_1, C_2) n \log n$, with high probability,

$$\boldsymbol{z}^*\nabla_{\boldsymbol{z}} f(\boldsymbol{z}) \ge \frac{199}{100}\left\|\boldsymbol{z}\right\|^4 - \frac{1001}{1000}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 - \left|\boldsymbol{x}^*\boldsymbol{z}\right|^2 \ge \frac{1}{500}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2$$

for all $\boldsymbol{z} \in \mathcal{R}_2^{\boldsymbol{z}}$, as desired. ■

17.4 Proof of Proposition 13.5

Proof We abbreviate $\phi(\boldsymbol{z})$ as ϕ below. Note that

$$(\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^*\nabla_{\boldsymbol{z}} f(\boldsymbol{z}) = \frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 (\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^*\boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z} - \frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2 (\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^*\boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z}.$$

We first bound the second term. By Lemma 17.3, when $m \ge C_1 n\log n$ for a sufficiently large constant C_1, with high probability, for all $\boldsymbol{z}\in\mathbb{C}^n$,

$$\begin{aligned}
&\Re\left(\frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2 (\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^*\boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z}\right)\\
&= \Re\left((\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^*(\left\|\boldsymbol{x}\right\|^2\boldsymbol{I}+\boldsymbol{x}\boldsymbol{x}^*)\boldsymbol{z}\right) + \Re\left((\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^*\boldsymbol{\Delta}\boldsymbol{z}\right) \quad (\text{for a certain } \boldsymbol{\Delta} \text{ with } \left\|\boldsymbol{\Delta}\right\| \le \left\|\boldsymbol{x}\right\|^2/100)\\
&\le \left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 + \left|\boldsymbol{x}^*\boldsymbol{z}\right|^2 - 2\left\|\boldsymbol{x}\right\|^2\left|\boldsymbol{x}^*\boldsymbol{z}\right| + \frac{1}{1000}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}\right\|\left\|\boldsymbol{z}\right\|.
\end{aligned}$$

To bound the first term, for a fixed τ to be determined later, define:

$$\begin{aligned}
S(\boldsymbol{z}) &\doteq \frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 \Re\left((\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^*\boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z}\right),\\
S_1(\boldsymbol{z}) &\doteq \frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 \Re\left((\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^*\boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z}\right)\mathbb{1}_{\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|\le\tau}\\
S_2(\boldsymbol{z}) &\doteq \frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 \Re\left((\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^*\boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z}\right)\mathbb{1}_{\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|\le\tau}\mathbb{1}_{\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|\le\tau}.
\end{aligned}$$

Obviously $S_1(\boldsymbol{z}) \ge S_2(\boldsymbol{z})$ for all $\boldsymbol{z}$ as

$$\begin{aligned}
S_1(\boldsymbol{z}) - S_2(\boldsymbol{z}) &= \frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 \Re\left((\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\theta})^*\boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z}\right)\mathbb{1}_{\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|\le\tau}\mathbb{1}_{\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|>\tau}\\
&\ge \frac{1}{m}\sum_{k=1}^m \left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2\left(\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 - \left|\boldsymbol{a}_k^*\boldsymbol{x}\right|\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|\right)\mathbb{1}_{\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|\le\tau}\mathbb{1}_{\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|>\tau} \ge 0.
\end{aligned}$$

Now for an $\varepsilon\in(0,\left\|\boldsymbol{x}\right\|)$ to be fixed later, consider an ε-net N_ε for the ball $\mathbb{CB}^n(\left\|\boldsymbol{x}\right\|)$, with $|N_\varepsilon| \le (3\left\|\boldsymbol{x}\right\|/\varepsilon)^{2n}$. On the complement of the event $\left\{\max_{k\in[m]}\left|\boldsymbol{a}_k^*\boldsymbol{x}\right| > \tau\right\}$, we have for any $t>0$ that

$$\begin{aligned}
&\mathbb{P}\left[S(\boldsymbol{z}) - \mathbb{E}\left[S(\boldsymbol{z})\right] < -t,\ \forall\ \boldsymbol{z}\in N_\varepsilon\right]\\
&\quad\le |N_\varepsilon|\,\mathbb{P}\left[S(\boldsymbol{z}) - \mathbb{E}\left[S(\boldsymbol{z})\right] < -t\right]\\
&\quad\le |N_\varepsilon|\,\mathbb{P}\left[\ S_1(\boldsymbol{z}) - \mathbb{E}\left[S_1(\boldsymbol{z})\right] < -t + \left|\mathbb{E}\left[S_1(\boldsymbol{z})\right] - \mathbb{E}\left[S(\boldsymbol{z})\right]\right|\ \right].
\end{aligned}$$

Because $S_1(\boldsymbol{z}) \ge S_2(\boldsymbol{z})$ as shown above,

$$\begin{aligned} &\mathbb{P}\left[\, S_1(\boldsymbol{z}) - \mathbb{E}\left[S_1(\boldsymbol{z})\right] < -t + \left|\mathbb{E}\left[S_1(\boldsymbol{z})\right] - \mathbb{E}\left[S(\boldsymbol{z})\right]\right| \,\right] \\ &\quad \le \mathbb{P}\left[\, S_2(\boldsymbol{z}) - \mathbb{E}\left[S_2(\boldsymbol{z})\right] < -t + \left|\mathbb{E}\left[S_1(\boldsymbol{z})\right] - \mathbb{E}\left[S(\boldsymbol{z})\right]\right| + \left|\mathbb{E}\left[S_1(\boldsymbol{z})\right] - \mathbb{E}\left[S_2(\boldsymbol{z})\right]\right| \,\right]. \end{aligned}$$

Thus, the unconditional probability can be bounded as

$$\begin{aligned} &\mathbb{P}\left[S(\boldsymbol{z}) - \mathbb{E}\left[S(\boldsymbol{z})\right] < -t, \ \forall\, \boldsymbol{z} \in N_\varepsilon\right] \\ &\le \ |N_\varepsilon|\, \mathbb{P}\left[\, S_2(\boldsymbol{z}) - \mathbb{E}\left[S_2(\boldsymbol{z})\right] < -t + \left|\mathbb{E}\left[S_1(\boldsymbol{z})\right] - \mathbb{E}\left[S(\boldsymbol{z})\right]\right| + \left|\mathbb{E}\left[S_1(\boldsymbol{z})\right] - \mathbb{E}\left[S_2(\boldsymbol{z})\right]\right| \,\right] \\ &\qquad + \mathbb{P}\left[\max_{k \in [m]} \left|\boldsymbol{a}_k^* \boldsymbol{x}\right| > \tau\right]. \end{aligned}$$

Taking $\tau = \sqrt{10 \log m}\, \|\boldsymbol{x}\|$, we obtain

$$\begin{aligned} \mathbb{P}\left[\max_{k \in [m]} \left|\boldsymbol{a}_k^* \boldsymbol{x}\right| > \tau\right] &\le m \exp\left(-\frac{10 \log m}{2}\right) \le m^{-4}, \\ \left|\mathbb{E}\left[S_1(\boldsymbol{z})\right] - \mathbb{E}\left[S(\boldsymbol{z})\right]\right| &\le \sqrt{\mathbb{E}\left[\left|\boldsymbol{a}^* \boldsymbol{z}\right|^6 \left|\boldsymbol{a}^* \left(\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi}\right)\right|^2\right]} \sqrt{\mathbb{P}\left[\left|\boldsymbol{a}^* \boldsymbol{x}\right| > \tau\right]} \le 4\sqrt{3} m^{-5/2} \left\|\boldsymbol{z}\right\|^3 \left\|\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi}\right\|, \\ \left|\mathbb{E}\left[S_1(\boldsymbol{z})\right] - \mathbb{E}\left[S_2(\boldsymbol{z})\right]\right| &\le \sqrt{\mathbb{E}\left[\left|\boldsymbol{a}^* \boldsymbol{z}\right|^6 \left|\boldsymbol{a}^* \left(\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi}\right)\right|^2 \mathbb{1}_{\left|\boldsymbol{a}^* \boldsymbol{x}\right| \le \tau}\right]} \sqrt{\mathbb{P}\left[\left|\boldsymbol{a}^* \boldsymbol{z}\right| > \tau\right]} \\ &\le 4\sqrt{3} m^{-5/2} \left\|\boldsymbol{z}\right\|^3 \left\|\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi}\right\|, \end{aligned}$$

where we have used $\|\boldsymbol{z}\| \le \|\boldsymbol{x}\|$ to simplify the last inequality. Now we used the moment-control Bernstein's inequality (Lemma A.1) to get a bound for probability on deviation of $S_2(\boldsymbol{z})$. To this end, we have

$$\begin{aligned} \mathbb{E}\left[\left|\boldsymbol{a}^* \boldsymbol{z}\right|^6 \left|\boldsymbol{a}^* (\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi})\right|^2 \mathbb{1}_{\left|\boldsymbol{a}^* \boldsymbol{x}\right| \le \tau} \mathbb{1}_{\left|\boldsymbol{a}^* \boldsymbol{z}\right| \le \tau}\right] &\le \tau^2 \mathbb{E}\left[\left|\boldsymbol{a}^* \boldsymbol{z}\right|^4 \left|\boldsymbol{a}^* (\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi})\right|^2\right] \\ &\le 240 \log m \left\|\boldsymbol{x}\right\|^2 \left\|\boldsymbol{z}\right\|^4 \left\|\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi}\right\|^2 \\ \mathbb{E}\left[\left|\boldsymbol{a}^* \boldsymbol{z}\right|^{3p} \left|\boldsymbol{a}^* (\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi})\right|^p \mathbb{1}_{\left|\boldsymbol{a}^* \boldsymbol{x}\right| \le \tau} \mathbb{1}_{\left|\boldsymbol{a}^* \boldsymbol{z}\right| \le \tau}\right] &\le \tau^{2p} \mathbb{E}\left[\left|\boldsymbol{a}^* \boldsymbol{z}\right|^p \left|\boldsymbol{a}^* (\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi})\right|^p\right] \\ &\le \left(10 \log m \left\|\boldsymbol{x}\right\|^2\right)^p p! \left\|\boldsymbol{z}\right\|^p \left\|\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi}\right\|^p, \end{aligned}$$

where the second inequality holds for any integer $p \ge 3$. Hence one can take

$$\begin{aligned} \sigma^2 &= 240 \log^2 m \left\|\boldsymbol{x}\right\|^4 \left\|\boldsymbol{z}\right\|^2 \left\|\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi}\right\|^2, \\ R &= 10 \log m \left\|\boldsymbol{x}\right\|^2 \left\|\boldsymbol{z}\right\| \left\|\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi}\right\| \end{aligned}$$

in Lemma A.1, and

$$t = \frac{1}{1000} \left\|\boldsymbol{x}\right\|^2 \left\|\boldsymbol{z}\right\| \left\|\boldsymbol{z} - \boldsymbol{x} \mathrm{e}^{\mathrm{i}\phi}\right\|$$

in the deviation inequality of $S_2(\boldsymbol{z})$ to obtain

$$\mathbb{P}\left[S_2(\boldsymbol{z}) - \mathbb{E}\left[S_2(\boldsymbol{z})\right] < -\frac{1}{200}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|\left\|\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}\right\|\right] \le \exp\left(-\frac{c_2 m}{\log^2 m}\right),$$

where we have used the fact $\|\boldsymbol{z}\| \le \|\boldsymbol{x}\|$ and assumed $4\sqrt{3}m^{-5/2} \le 1/200$ to simplify the probability. Thus, with probability at least $1 - m^{-4} - \exp\left(-c_2 m/\log^2 m + 2n\log(3\left\|\boldsymbol{x}\right\|/\varepsilon)\right)$, it holds that

$$S(\boldsymbol{z}) \ge 2\left\|\boldsymbol{z}\right\|^4 - 2\left\|\boldsymbol{z}\right\|^2\left|\boldsymbol{x}^*\boldsymbol{z}\right| - \frac{1}{1000}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|\left\|\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}\right\| \quad \forall\ \boldsymbol{z}\in N_\varepsilon. \tag{17.4.1}$$

Moreover, for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{R}_2^{\boldsymbol{h}}$, we have

$$\begin{aligned}
&\left|S(\boldsymbol{z}) - S(\boldsymbol{z}')\right| \\
&\le \frac{1}{m}\sum_{k=1}^m \left|\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 - \left|\boldsymbol{a}_k^*\boldsymbol{z}'\right|^2\right|\left|\boldsymbol{h}^*(\boldsymbol{z})\boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z}\right| + \frac{1}{m}\sum_{k=1}^m\left|\boldsymbol{a}_k^*\boldsymbol{z}'\right|^2\left|\boldsymbol{h}^*(\boldsymbol{z})\boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z} - \boldsymbol{h}^*(\boldsymbol{z}')\boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z}'\right| \\
&\le 4\max_{k\in[m]}\left\|\boldsymbol{a}_k\right\|^4\left\|\boldsymbol{x}\right\|^3\left\|\boldsymbol{z}-\boldsymbol{z}'\right\| + 5\max_{k\in[m]}\left\|\boldsymbol{a}_k\right\|^4\left\|\boldsymbol{x}\right\|^3\left\|\boldsymbol{z}-\boldsymbol{z}'\right\| \\
&\le 90n^2\log^2 m\left\|\boldsymbol{x}\right\|^3\left\|\boldsymbol{z}-\boldsymbol{z}'\right\|,
\end{aligned}$$

as $\max_{k\in[m]}\left\|\boldsymbol{a}_k\right\|^4 \le 10n^2\log^2 m$ with probability at least $1 - c_3 m^{-n}$, and $11\left\|\boldsymbol{x}\right\|/20 \le \|\boldsymbol{z}\| \le \|\boldsymbol{x}\|$, and also $\|\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z})} - \boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi(\boldsymbol{z}')}\| \le 2\left\|\boldsymbol{z}-\boldsymbol{z}'\right\|$ for $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{R}_2^{\boldsymbol{h}}$. Every $\boldsymbol{z}\in\mathcal{R}_2^{\boldsymbol{h}}$ can be written as $\boldsymbol{z} = \boldsymbol{z}' + \boldsymbol{e}$, with $\boldsymbol{z}' \in N_\varepsilon$ and $\|\boldsymbol{e}\| \le \varepsilon$. Thus,

$$\begin{aligned}
S(\boldsymbol{z}) &\ge S(\boldsymbol{z}') - 90n^2\log^2 m\left\|\boldsymbol{x}\right\|^3\varepsilon \\
&\ge 2\left\|\boldsymbol{z}'\right\|^4 - 2\left\|\boldsymbol{z}'\right\|^2\left|\boldsymbol{x}^*\boldsymbol{z}'\right| - \frac{1}{1000}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}'\right\|\left\|\boldsymbol{z}'-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}\right\| - 90n^2\log^2 m\left\|\boldsymbol{x}\right\|^3\varepsilon \\
&\ge 2\left\|\boldsymbol{z}\right\|^4 - 2\left\|\boldsymbol{z}\right\|^2\left|\boldsymbol{x}^*\boldsymbol{z}\right| - \frac{1}{1000}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|\left\|\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}\right\| - 11\varepsilon\left\|\boldsymbol{x}\right\|^3 - 90n^2\log^2 m\left\|\boldsymbol{x}\right\|^3\varepsilon,
\end{aligned}$$

where the additional $11\varepsilon\left\|\boldsymbol{x}\right\|^3$ term in the third line is to account for the change from $\boldsymbol{z}'$ to $\boldsymbol{z}$, which has been simplified by assumptions that $11/20\cdot\|\boldsymbol{x}\| \le \|\boldsymbol{z}\| \le \|\boldsymbol{x}\|$ and that $\varepsilon \le \|\boldsymbol{x}\|$. Choosing $\varepsilon = \left\|\boldsymbol{x}\right\|/(c_5 n^2\log^2 m)$ for a sufficiently large $c_5 > 0$ and additionally using $\operatorname{dist}(\boldsymbol{z}, X) \ge \left\|\boldsymbol{x}\right\|/3$, we obtain that

$$S(\boldsymbol{z}) \ge 2\left\|\boldsymbol{z}\right\|^4 - 2\left\|\boldsymbol{z}\right\|^2\Re\left(\boldsymbol{x}^*\boldsymbol{z}\mathrm{e}^{-\mathrm{i}\phi}\right) - \frac{1}{500}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|\left\|\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}\right\| \tag{17.4.2}$$

for all $\boldsymbol{z}\in\mathcal{R}_2^{\boldsymbol{h}}$, with probability at least $1 - c_6 m^{-1} - c_7\exp\left(-c_2 m/\log^2 m + c_9 n\log(C_8 n\log m)\right)$.

Combining the above estimates, when $m \ge C_{10} n\log^3 n$ for sufficiently large constant C_{10}, with high probability,

$$\Re\left((\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi})^*\nabla_{\boldsymbol{z}} f(\boldsymbol{z})\right) \ge \frac{1}{1000}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|\left\|\boldsymbol{z}-\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}\right\|$$

for all $\boldsymbol{z} \in \mathcal{R}_2^{\boldsymbol{h}}$, as claimed. ■

17.5 Proof of Proposition 13.6

Proof It is enough to prove that for all unit vectors $\boldsymbol{g}$ that are geometrically orthogonal to i$\boldsymbol{x}$, i.e., $\boldsymbol{g} \in \mathcal{T} \doteq \{\boldsymbol{z} : \Im(\boldsymbol{z}^*\boldsymbol{x}) = 0, \|\boldsymbol{z}\| = 1\}$ and all $t \in [0, \|\boldsymbol{x}\|/\sqrt{7}]$, the following holds:

$$\begin{bmatrix}\boldsymbol{g}\\ \overline{\boldsymbol{g}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{x}+t\boldsymbol{g}) \begin{bmatrix}\boldsymbol{g}\\ \overline{\boldsymbol{g}}\end{bmatrix} \ge \frac{1}{4}\left\|\boldsymbol{x}\right\|^2.$$

Direct calculation shows

$$\begin{aligned}
&\begin{bmatrix}\boldsymbol{g}\\ \overline{\boldsymbol{g}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{x}+t\boldsymbol{g}) \begin{bmatrix}\boldsymbol{g}\\ \overline{\boldsymbol{g}}\end{bmatrix}\\
&= \frac{1}{m}\sum_{k=1}^m 4\left|\boldsymbol{a}_k^*(\boldsymbol{x}+t\boldsymbol{g})\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{g}\right|^2 - 2\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{g}\right|^2 + 2\Re\left[(t\boldsymbol{a}_k^*\boldsymbol{g}+\boldsymbol{a}_k^*\boldsymbol{x})^2(\boldsymbol{g}^*\boldsymbol{a}_k)^2\right]\\
&\ge \frac{1}{m}\sum_{k=1}^m 4\left|\boldsymbol{a}_k^*(\boldsymbol{x}+t\boldsymbol{g})\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{g}\right|^2 - 2\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{g}\right|^2 + 4\left[\Re(t\boldsymbol{a}_k^*\boldsymbol{g}+\boldsymbol{a}_k^*\boldsymbol{x})(\boldsymbol{g}^*\boldsymbol{a}_k)\right]^2 - 2\left|(t\boldsymbol{a}_k^*\boldsymbol{g}+\boldsymbol{a}_k^*\boldsymbol{x})(\boldsymbol{g}^*\boldsymbol{a}_k)\right|^2\\
&\ge \frac{1}{m}\sum_{k=1}^m 2\left|\boldsymbol{a}_k^*(\boldsymbol{x}+t\boldsymbol{g})\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{g}\right|^2 - 2\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{g}\right|^2 + 4\left[\Re(t\boldsymbol{a}_k^*\boldsymbol{g}+\boldsymbol{a}_k^*\boldsymbol{x})(\boldsymbol{g}^*\boldsymbol{a}_k)\right]^2.
\end{aligned}$$

Lemma 17.4 implies when $m \ge C_1 n\log n$ for sufficiently large constant C_1, with high probability,

$$\frac{1}{m}\sum_{k=1}^m 2\left|\boldsymbol{a}_k^*(\boldsymbol{x}+t\boldsymbol{g})\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{g}\right|^2 \ge \frac{199}{100}\left|(\boldsymbol{x}+t\boldsymbol{g})^*\boldsymbol{g}\right|^2 + \frac{199}{100}\left\|\boldsymbol{x}+t\boldsymbol{g}\right\|^2\left\|\boldsymbol{g}\right\|^2 \qquad (17.5.1)$$

for all $\boldsymbol{g} \in \mathbb{C}^n$ and all $t \in [0, \|\boldsymbol{x}\|/\sqrt{7}]$. Lemma 17.3 implies that when $m \ge C_2 n\log n$ for sufficiently large constant C_2, with high probability,

$$\frac{1}{m}\sum_{k=1}^m 2\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2\left|\boldsymbol{a}_k^*\boldsymbol{g}\right|^2 \le \frac{201}{100}\left|\boldsymbol{x}^*\boldsymbol{g}\right|^2 + \frac{201}{100}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{g}\right\|^2 \qquad (17.5.2)$$

for all $\boldsymbol{g} \in \mathbb{C}^n$. Moreover, Lemma 17.4 implies when $m \geq C_3 n \log n$ for sufficiently large constant C_3, with high probability,

$$\frac{4}{m}\sum_{k=1}^{m}\left[\Re(t\boldsymbol{a}_k^*\boldsymbol{g} + \boldsymbol{a}_k^*\boldsymbol{x})(\boldsymbol{g}^*\boldsymbol{a}_k)\right]^2 \geq 2\left\|\boldsymbol{x}+t\boldsymbol{g}\right\|^2\left\|\boldsymbol{g}\right\|^2 + 6\left|(\boldsymbol{x}+\boldsymbol{g})^*\boldsymbol{g}\right|^2 - \frac{1}{400}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{g}\right\|^2$$

for all $\boldsymbol{g} \in \mathcal{T}$, where we have used that $\Im(\boldsymbol{g}^*\boldsymbol{x}) = 0 \Longrightarrow \Im(\boldsymbol{x}+\boldsymbol{g})^*\boldsymbol{g} = 0$ to simplify the results.

Collecting the above estimates, we obtain that when $m \geq C_4 n \log n$, with high probability,

$$\begin{aligned}
&\begin{bmatrix}\boldsymbol{g}\\ \overline{\boldsymbol{g}}\end{bmatrix}^*\nabla^2 f(\boldsymbol{x}+t\boldsymbol{g})\begin{bmatrix}\boldsymbol{g}\\ \overline{\boldsymbol{g}}\end{bmatrix}\\
\geq\ & \left(\frac{399}{100}\left\|\boldsymbol{x}+t\boldsymbol{g}\right\|^2 - \frac{161}{80}\left\|\boldsymbol{x}\right\|^2\right) + \left(\frac{799}{100}\left|(\boldsymbol{x}+t\boldsymbol{g})^*\boldsymbol{g}\right|^2 - \frac{201}{100}\left|\boldsymbol{x}^*\boldsymbol{g}\right|^2\right)\\
=\ & \frac{791}{400}\left\|\boldsymbol{x}\right\|^2 + \frac{598}{100}\left|\boldsymbol{x}^*\boldsymbol{g}\right|^2 + \frac{1198}{100}t^2 + \frac{2396}{100}t\Re(\boldsymbol{x}^*\boldsymbol{g}).
\end{aligned}$$

To provide a lower bound for the above, we let $\Re(\boldsymbol{x}^*\boldsymbol{g}) = \boldsymbol{x}^*\boldsymbol{g} = \lambda\left\|\boldsymbol{x}\right\|$ with $\lambda \in [-1,1]$ and $t = \eta\left\|\boldsymbol{x}\right\|$ with $\eta \in [0, 1/\sqrt{7}]$. Then

$$\frac{598}{100}\left|\boldsymbol{x}^*\boldsymbol{g}\right|^2 + \frac{1198}{100}t^2 + \frac{2396}{100}t\Re(\boldsymbol{x}^*\boldsymbol{g}) = \left\|\boldsymbol{x}\right\|^2\left(\frac{598}{100}\lambda^2 + \frac{1198}{100}\eta^2 + \frac{2396}{100}\lambda\eta\right) \doteq \left\|\boldsymbol{x}\right\|^2\phi(\lambda,\eta).$$

For any fixed η, it is easy to see that minimizer occurs when $\lambda = -\frac{599}{299}\eta$. Plugging this into $\phi(\lambda,\eta)$, one obtains $\phi(\lambda,\eta) \geq -\frac{241}{20}\eta^2 \geq -\frac{241}{140}$. Thus,

$$\begin{bmatrix}\boldsymbol{g}\\ \overline{\boldsymbol{g}}\end{bmatrix}^*\nabla^2 f(\boldsymbol{x}+t\boldsymbol{g})\begin{bmatrix}\boldsymbol{g}\\ \overline{\boldsymbol{g}}\end{bmatrix} \geq \left(\frac{791}{400} - \frac{241}{140}\right)\left\|\boldsymbol{x}\right\|^2 \geq \frac{1}{4}\left\|\boldsymbol{x}\right\|^2,$$

as claimed. ■

17.6 Proof of Proposition 13.7

Proof For convenience, we will define a relaxed $\mathcal{R}_2^{\boldsymbol{h}}$ region

$$\mathcal{R}_2^{\boldsymbol{h}'} \doteq \left\{\boldsymbol{z} : \Re\left(\langle \boldsymbol{h}(\boldsymbol{z}), \nabla_{\boldsymbol{z}}\mathbb{E}\left[f\right]\rangle\right) \geq \frac{1}{250}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|\left\|\boldsymbol{h}(\boldsymbol{z})\right\|, \left\|\boldsymbol{z}\right\| \leq \left\|\boldsymbol{x}\right\|\right\} \supset \mathcal{R}_2^{\boldsymbol{h}}$$

and try to show that $\mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}} \cup \mathcal{R}_2^{\boldsymbol{h}'} \cup \mathcal{R}_3 = \mathbb{C}^n$. In the end, we will discuss how this implies the claimed result.

We will first divide $\mathbb{C}^n$ into several (overlapping) regions, and show that each such region is a subset of

$\mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}} \cup \mathcal{R}_2^{\boldsymbol{h}'} \cup \mathcal{R}_3$.

Cover $\mathcal{R}_a \doteq \left\{\boldsymbol{z} : |\boldsymbol{x}^* \boldsymbol{z}| \le \frac{1}{2} \|\boldsymbol{x}\| \|\boldsymbol{z}\|\right\}$: In this case, when $\|\boldsymbol{z}\|^2 \le \frac{398}{601} \|\boldsymbol{x}\|^2$,

$$8 |\boldsymbol{x}^* \boldsymbol{z}|^2 + \frac{401}{100} \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 \le \frac{601}{100} \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 \le \frac{398}{100} \|\boldsymbol{x}\|^4 .$$

On the other hand, when $\|\boldsymbol{z}\|^2 \ge \frac{626}{995} \|\boldsymbol{x}\|^2$,

$$\frac{501}{500} \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 + |\boldsymbol{x}^* \boldsymbol{z}|^2 \le \frac{313}{250} \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 \le \frac{199}{100} \|\boldsymbol{z}\|^4 .$$

Since $\frac{398}{601} > \frac{626}{995}$, we conclude that $\mathcal{R}_a \subset \mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}}$.

Cover $\mathcal{R}_b \doteq \left\{\boldsymbol{z} : \ |\boldsymbol{x}^* \boldsymbol{z}| \ge \frac{1}{2} \|\boldsymbol{x}\| \|\boldsymbol{z}\| \, , \ \|\boldsymbol{z}\| \le \frac{57}{100} \|\boldsymbol{x}\|\right\}$: In this case,

$$8 |\boldsymbol{x}^* \boldsymbol{z}|^2 + \frac{401}{100} \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 \le \frac{1201}{100} \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 \le \frac{398}{100} \|\boldsymbol{x}\|^4 .$$

So $\mathcal{R}_b$ is covered by $\mathcal{R}_1$.

Cover $\mathcal{R}_c \doteq \left\{\boldsymbol{z} : \frac{1}{2} \|\boldsymbol{x}\| \|\boldsymbol{z}\| \le |\boldsymbol{x}^* \boldsymbol{z}| \le \frac{99}{100} \|\boldsymbol{x}\| \|\boldsymbol{z}\| \, , \ \|\boldsymbol{z}\| \ge \frac{11}{20} \|\boldsymbol{x}\|\right\}$: We show this region is covered by $\mathcal{R}_2^{\boldsymbol{z}}$ and $\mathcal{R}_2^{\boldsymbol{h}'}$. First, for any $\boldsymbol{z} \in \mathcal{R}_c$, when $\|\boldsymbol{z}\| \ge \sqrt{\frac{1983}{1990}} \|\boldsymbol{x}\|$,

$$\frac{501}{500} \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 + |\boldsymbol{x}^* \boldsymbol{z}|^2 \le \frac{1983}{1000} \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 \le \frac{199}{100} \|\boldsymbol{z}\|^4 ,$$

implying that $\mathcal{R}_c \cap \left\{\boldsymbol{z} : \|\boldsymbol{z}\| \ge \sqrt{\frac{1983}{1990}} \|\boldsymbol{x}\|\right\} \subset \mathcal{R}_2^{\boldsymbol{z}}$. Next we suppose $\|\boldsymbol{z}\| = \lambda \|\boldsymbol{x}\|$ and $|\boldsymbol{x}^* \boldsymbol{z}| = \eta \|\boldsymbol{x}\| \|\boldsymbol{z}\|$, where $\lambda \in [\frac{11}{20}, \sqrt{\frac{1984}{1990}}]$ and $\eta \in [\frac{1}{2}, \frac{99}{100}]$, and show the rest of $\mathcal{R}_c$ is covered by $\mathcal{R}_2^{\boldsymbol{h}'}$. To this end, it is enough to verify that

$$2 \left(\|\boldsymbol{x}\|^2 - \|\boldsymbol{z}\|^2\right) |\boldsymbol{x}^* \boldsymbol{z}| + 2 \|\boldsymbol{z}\|^4 - \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\|^2 - |\boldsymbol{x}^* \boldsymbol{z}|^2 - \frac{1}{250} \|\boldsymbol{x}\|^2 \|\boldsymbol{z}\| \sqrt{\|\boldsymbol{x}\|^2 + \|\boldsymbol{z}\|^2 - 2 |\boldsymbol{x}^* \boldsymbol{z}|} \ge 0$$

over this subregion. Writing the left as a function of λ, η and eliminating $\|\boldsymbol{x}\|$ and $\|\boldsymbol{z}\|$, it is enough to show

$$h(\lambda, \eta) \doteq 2(1 - \lambda^2)\eta + 2\lambda^3 - \lambda - \eta^2 \lambda - \frac{1}{250} \sqrt{1 + \lambda^2 - 2\eta\lambda} \ge 0,$$

which is implied by

$$p(\lambda, \eta) \doteq 2(1 - \lambda^2)\eta + 2\lambda^3 - \lambda - \eta^2 \lambda \ge \frac{49}{10000},$$

as $\frac{1}{250} \sqrt{1 + \lambda^2 - 2\eta\lambda} < 49/10000$. Let $\boldsymbol{H}_p$ be the Hessian matrix of this bivariate function, it is easy to verify that $\det(\boldsymbol{H}_p) = -4(\eta + \lambda)^2 - 36\lambda^2 < 0$ for all valid (λ, η). Thus, the minimizer must occur on the boundary. For any fixed λ, $2(1-\lambda^2)\eta-\eta^2\lambda$ is minimized at either $\eta = 99/100$ or $\eta = 1/2$. When $\eta = 99/100$, p is minimized at $\lambda = (4\cdot 0.99+\sqrt{40\cdot 0.99^2+24})/12 < \sqrt{1984/1990}$, giving $p \ge 0.019$; when $\eta = 1/2$, p is minimized when $\lambda = (4\cdot 0.5+\sqrt{40\cdot 0.5^2+24}/12) = (2+\sqrt{34})/12$, giving $p \ge 0.3$. Overall, $p \ge 0.019 > 49/10000$, as desired.

Cover $\mathcal{R}_d \doteq \left\{\boldsymbol{z} : \frac{99}{100}\left\|\boldsymbol{x}\right\|\left\|\boldsymbol{z}\right\| \le \left|\boldsymbol{x}^*\boldsymbol{z}\right| \le \left\|\boldsymbol{x}\right\|\left\|\boldsymbol{z}\right\|,\ \left\|\boldsymbol{z}\right\| \ge \frac{11}{20}\left\|\boldsymbol{x}\right\|\right\}$: We show that this region is covered by $\mathcal{R}_2^{\boldsymbol{z}}$, $\mathcal{R}_3$, and $\mathcal{R}_2^{\boldsymbol{h}'}$ together. First, for any $\boldsymbol{z} \in \mathcal{R}_d$, when $\left\|\boldsymbol{z}\right\| \ge \sqrt{\frac{1001}{995}}\left\|\boldsymbol{x}\right\|$,

$$\frac{501}{500}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 + \left|\boldsymbol{x}^*\boldsymbol{z}\right|^2 \le \frac{1001}{500}\left\|\boldsymbol{x}\right\|^2\left\|\boldsymbol{z}\right\|^2 \le \frac{199}{100}\left\|\boldsymbol{z}\right\|^4 .$$

So $\mathcal{R}_d \cap \left\{\boldsymbol{z} : \left\|\boldsymbol{z}\right\| \ge \sqrt{\frac{1001}{995}}\left\|\boldsymbol{x}\right\|\right\} \subset \mathcal{R}_2^{\boldsymbol{z}}$. Next, we show that any $\boldsymbol{z} \in \mathcal{R}_d$ with $\left\|\boldsymbol{z}\right\| \le 24/25 \cdot \left\|\boldsymbol{x}\right\|$ is contained in $\mathcal{R}_2^{\boldsymbol{h}'}$. Similar to the above argument for $\mathcal{R}_c$, it is enough to show

$$p(\lambda,\eta) \doteq 2(1-\lambda^2)\eta + 2\lambda^3 - \lambda - \eta^2\lambda \ge 0.00185,$$

as $\frac{1}{250}\sqrt{1+\lambda^2-2\eta\lambda} < 0.00185$ in this case. Since the Hessian is again always indefinite, we check the optimal value for $\eta = 99/100$ and $\eta = 1$ and do the comparison. It can be verified $p \ge 0.00627 > 0.00185$ in this case. So $\mathcal{R}_d \cap \left\{\boldsymbol{z} : \left\|\boldsymbol{z}\right\| \le \frac{24}{25}\left\|\boldsymbol{x}\right\|\right\} \subset \mathcal{R}_2^{\boldsymbol{h}'}$. Finally, we consider the case $\frac{23}{25}\left\|\boldsymbol{x}\right\| \le \left\|\boldsymbol{z}\right\| \le \sqrt{\frac{1005}{995}}\left\|\boldsymbol{x}\right\|$. A λ, η argument as above leads to

$$\left\|\boldsymbol{h}(\boldsymbol{z})\right\|^2 = \left\|\boldsymbol{x}\right\|^2 + \left\|\boldsymbol{z}\right\|^2 - 2\left|\boldsymbol{x}^*\boldsymbol{z}\right| < \frac{1}{7}\left\|\boldsymbol{x}\right\|^2,$$

implying that $\mathcal{R}_d \cap \left\{\boldsymbol{z} : \frac{23}{25}\left\|\boldsymbol{x}\right\| \le \left\|\boldsymbol{z}\right\| \le \sqrt{\frac{1005}{995}}\left\|\boldsymbol{x}\right\|\right\} \subset \mathcal{R}_3$.

In summary, now we obtain that $\mathbb{C}^n = \mathcal{R}_a \cup \mathcal{R}_b \cup \mathcal{R}_c \cup \mathcal{R}_d \subset \mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}} \cup \mathcal{R}_2^{\boldsymbol{h}'} \cup \mathcal{R}_3$. Observe that $\mathcal{R}_{\boldsymbol{h}'}$ is only used to cover $\mathcal{R}_c \cup \mathcal{R}_d$, which is in turn a subset of $\{\boldsymbol{z} : \left\|\boldsymbol{z}\right\| \ge 11\left\|\boldsymbol{x}\right\|/20\}$. Thus, $\mathbb{C}^n = \mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}} \cup (\mathcal{R}_2^{\boldsymbol{h}'} \cap \{\boldsymbol{z} : \left\|\boldsymbol{z}\right\| \ge 11\left\|\boldsymbol{x}\right\|/20\}) \cup \mathcal{R}_3$. Moreover, by the definition of $\mathcal{R}_3$,

$$\begin{aligned}
&\mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}} \cup (\mathcal{R}_2^{\boldsymbol{h}'} \cap \{\boldsymbol{z} : \left\|\boldsymbol{z}\right\| \ge 11\left\|\boldsymbol{x}\right\|/20\}) \cup \mathcal{R}_3 \\
&\subset \mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}} \cup (\mathcal{R}_2^{\boldsymbol{h}'} \cap \{\boldsymbol{z} : \left\|\boldsymbol{z}\right\| \ge 11\left\|\boldsymbol{x}\right\|/20\} \cap \mathcal{R}_3^c) \cup \mathcal{R}_3 \\
&\subset \mathcal{R}_1 \cup \mathcal{R}_2^{\boldsymbol{z}} \cup \mathcal{R}_2^{\boldsymbol{h}} \cup \mathcal{R}_3 \subset \mathbb{C}^n,
\end{aligned}$$

implying the claimed coverage. ■

Chapter 18

Proofs of Technical Results for Trust-Region Algorithm

> I have tried to avoid long numerical computations, thereby following Riemann's postulate that proofs should be given through ideas and not voluminous computations.
>
> David Hilbert

18.1 Auxiliary lemmas

Lemma 18.1 *When $m \geq Cn$ for a sufficiently large C, it holds with probability at least $1 - c_a \exp(-c_b m)$ that*

$$\frac{1}{m}\sum_{k=1}^{m}\left|\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 - \left|\boldsymbol{a}_k^*\boldsymbol{w}\right|^2\right| \leq \frac{3}{2}\left\|\boldsymbol{z}-\boldsymbol{w}\right\|\left(\left\|\boldsymbol{z}\right\| + \left\|\boldsymbol{w}\right\|\right)$$

for all $\boldsymbol{z}, \boldsymbol{w} \in \mathbb{C}^n$. Here C, c_a, c_b are positive absolute constants.

Proof Lemma 3.1 in [CSV13] has shown that when $m \geq C_1 n$, it holds with probability at least $1 - c_2 \exp(-c_3 m)$ that

$$\frac{1}{m}\sum_{k=1}^{m}\left|\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 - \left|\boldsymbol{a}_k^*\boldsymbol{w}\right|^2\right| \leq \frac{3}{2\sqrt{2}}\left\|\boldsymbol{z}\boldsymbol{z}^* - \boldsymbol{w}\boldsymbol{w}^*\right\|_*$$

for all $\boldsymbol{z}$ and $\boldsymbol{w}$, where $\left\|\cdot\right\|_*$ is the nuclear norm that sums up singular values. The claims follows from

$$\left\|\boldsymbol{z}\boldsymbol{z}^* - \boldsymbol{w}\boldsymbol{w}^*\right\|_* \leq \sqrt{2}\left\|\boldsymbol{z}\boldsymbol{z}^* - \boldsymbol{w}\boldsymbol{w}^*\right\| \leq \sqrt{2}\left\|\boldsymbol{z}-\boldsymbol{w}\right\|\left(\left\|\boldsymbol{z}\right\| + \left\|\boldsymbol{w}\right\|\right),$$

completing the proof. ■

Lemma 18.2 *When $m \geq Cn \log n$, with probability at least $1 - c_a m^{-1} - c_b \exp\left(-c_c m / \log m\right)$,*

$$\left\| \nabla^2 f(\boldsymbol{x} \mathrm{e}^{\mathrm{i}\psi}) - \mathbb{E}\left[\nabla^2 f(\boldsymbol{x} \mathrm{e}^{\mathrm{i}\psi})\right] \right\| \leq \frac{1}{100} \left\| \boldsymbol{x} \right\|^2$$

for all $\psi \in [0, 2\pi)$. Here C, c_a to c_c are positive absolute constants.

Proof By Lemma 17.3, we have that

$$\begin{aligned} & \left\| \nabla^2 f(\boldsymbol{x} \mathrm{e}^{\mathrm{i}\psi}) - \mathbb{E}\left[\nabla^2 f(\boldsymbol{x} \mathrm{e}^{\mathrm{i}\psi})\right] \right\| \\ \leq & \left\| \frac{1}{m} \sum_{k=1}^m \left| \boldsymbol{a}_k^* \boldsymbol{x} \right|^2 \boldsymbol{a}_k \boldsymbol{a}_k - \left(\left\| \boldsymbol{x} \right\|^2 \boldsymbol{I} + \boldsymbol{x} \boldsymbol{x}^* \right) \right\| + \left\| \frac{1}{m} \sum_{k=1}^m (\boldsymbol{a}_k^* \boldsymbol{x})^2 \boldsymbol{a}_k \boldsymbol{a}_k^\top \mathrm{e}^{\mathrm{i}2\psi} - 2 \boldsymbol{x} \boldsymbol{x}^\top \mathrm{e}^{\mathrm{i}2\psi} \right\| \\ \leq & \frac{1}{200} \left\| \boldsymbol{x} \right\|^2 + \frac{1}{200} \left\| \boldsymbol{x} \right\|^2 \leq \frac{1}{100} \left\| \boldsymbol{x} \right\|^2 \end{aligned}$$

holds with high probability when $m \geq C_1 n \log n$ for a sufficiently large C_1. ■

18.2 Proof of Lemma 14.1

Proof For any $\boldsymbol{z}, \boldsymbol{z}' \in \Gamma'$, we have

$$\begin{aligned} \left| f(\boldsymbol{z}) - f(\boldsymbol{z}') \right| &= \frac{1}{2m} \left| \sum_{k=1}^m \left| \boldsymbol{a}_k^* \boldsymbol{z} \right|^4 - \left| \boldsymbol{a}_k^* \boldsymbol{z}' \right|^4 - 2 \sum_{k=1}^m \left| \boldsymbol{a}_k^* \boldsymbol{x} \right|^2 \left(\left| \boldsymbol{a}_k^* \boldsymbol{z} \right|^2 - \left| \boldsymbol{a}_k^* \boldsymbol{z}' \right|^2 \right) \right| \\ &\leq \frac{1}{2m} \sum_{k=1}^m (\left| \boldsymbol{a}_k^* \boldsymbol{z} \right|^2 + \left| \boldsymbol{a}_k^* \boldsymbol{z}' \right|^2) \left| \left| \boldsymbol{a}_k^* \boldsymbol{z} \right|^2 - \left| \boldsymbol{a}_k^* \boldsymbol{z}' \right|^2 \right| + \frac{1}{m} \sum_{k=1}^m \left| \boldsymbol{a}_k^* \boldsymbol{x} \right|^2 \left| \left| \boldsymbol{a}_k^* \boldsymbol{z} \right|^2 - \left| \boldsymbol{a}_k^* \boldsymbol{z}' \right|^2 \right| \\ &\leq 4 R_1^2 \left\| \boldsymbol{A} \right\|_{\ell^1 \to \ell^2}^2 \cdot \frac{3}{2} \cdot 4 R_1 \left\| \boldsymbol{z} - \boldsymbol{z}' \right\| + 2 \left\| \boldsymbol{A} \right\|_{\ell^1 \to \ell^2}^2 \left\| \boldsymbol{x} \right\|^2 \cdot \frac{3}{2} \cdot 4 R_1 \left\| \boldsymbol{z} - \boldsymbol{z}' \right\| \\ &\leq (24 R_1^3 \left\| \boldsymbol{A} \right\|_{\ell^1 \to \ell^2}^2 + 12 \left\| \boldsymbol{A} \right\|_{\ell^1 \to \ell^2}^2 \left\| \boldsymbol{x} \right\|^2 R_1) \left\| \boldsymbol{z} - \boldsymbol{z}' \right\|, \end{aligned}$$

where in the third line we invoked results of Lemma 18.1, and hence the derived inequality holds with high probability when $m \geq C_1 n$. Similarly, for the gradient,

$$\begin{aligned} & \left\| \nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{z}') \right\| \\ = & \frac{\sqrt{2}}{m} \left\| \sum_{k=1}^m \left(\left| \boldsymbol{a}_k^* \boldsymbol{z} \right|^2 - \left| \boldsymbol{a}_k^* \boldsymbol{x} \right|^2 \right) \boldsymbol{a}_k \boldsymbol{a}_k^* \boldsymbol{z} - \sum_{k=1}^m \left(\left| \boldsymbol{a}_k^* \boldsymbol{z}' \right|^2 - \left| \boldsymbol{a}_k^* \boldsymbol{x} \right|^2 \right) \boldsymbol{a}_k \boldsymbol{a}_k^* \boldsymbol{z}' \right\| \\ \leq & \frac{\sqrt{2}}{m} \sum_{k=1}^m \left\| (\left| \boldsymbol{a}_k^* \boldsymbol{z} \right|^2 - \left| \boldsymbol{a}_k^* \boldsymbol{z}' \right|^2) \boldsymbol{a}_k \boldsymbol{a}_k^* \boldsymbol{z} \right\| + \sqrt{2} \left\| \frac{1}{m} \sum_{k=1}^m \boldsymbol{a}_k \boldsymbol{a}_k^* \left| \boldsymbol{a}_k^* \boldsymbol{z}' \right|^2 \right\| \left\| \boldsymbol{z} - \boldsymbol{z}' \right\| \end{aligned}$$

$$
\begin{aligned}
&+ \sqrt{2}\left\|\frac{1}{m}\sum_{k=1}^{m}\boldsymbol{a}_k\boldsymbol{a}_k^*\left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2\right\|\left\|\boldsymbol{z}-\boldsymbol{z}'\right\| \\
\le\ & \sqrt{2}\left\|\boldsymbol{A}\right\|_{\ell^1\to\ell^2}^2\cdot 2R_1\cdot\frac{3}{2}\cdot 4R_1\left\|\boldsymbol{z}-\boldsymbol{z}'\right\| + (8\sqrt{2}\left\|\boldsymbol{A}\right\|_{\ell^1\to\ell^2}^2 R_1^2 + 2\sqrt{2}\left\|\boldsymbol{A}\right\|_{\ell^1\to\ell^2}^2\left\|\boldsymbol{x}\right\|^2)\left\|\boldsymbol{z}-\boldsymbol{z}'\right\| \\
\le\ & (20\sqrt{2}\left\|\boldsymbol{A}\right\|_{\ell^1\to\ell^2}^2 R_1^2 + 2\sqrt{2}\left\|\boldsymbol{A}\right\|_{\ell^1\to\ell^2}^2\left\|\boldsymbol{x}\right\|^2)\left\|\boldsymbol{z}-\boldsymbol{z}'\right\|,
\end{aligned}
$$

where from the second to the third inequality we used the fact $\left\|\frac{1}{m}\sum_{k=1}^{m}\boldsymbol{a}_k\boldsymbol{a}_k^*\right\| \le 2$ with probability at least $1-\exp(-c_2 m)$. Similarly for the Hessian,

$$
\begin{aligned}
&\left\|\nabla^2 f(\boldsymbol{z}) - \nabla^2 f(\boldsymbol{z}')\right\| \\
=\ & \sup_{\|\boldsymbol{w}\|=1}\left|\frac{1}{2}\begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix}^*\left(\nabla^2 f(\boldsymbol{z}) - \nabla^2 f(\boldsymbol{z}')\right)\begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix}\right| \\
\le\ & \sup_{\|\boldsymbol{w}\|=1} 2\left\|\frac{1}{m}\sum_{k=1}^{m}(\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 - \left|\boldsymbol{a}_k^*\boldsymbol{z}'\right|^2)\left|\boldsymbol{a}_k^*\boldsymbol{w}\right|^2\right\| + \left\|\frac{1}{m}\sum_{k=1}^{m}\Re((\boldsymbol{a}_k^*\boldsymbol{z})^2 - (\boldsymbol{a}_k^*\boldsymbol{z}')^2)(\boldsymbol{w}^*\boldsymbol{a}_k)^2\right\| \\
\le\ & 2\left\|\boldsymbol{A}\right\|_{\ell^1\to\ell^2}^2\cdot\frac{3}{2}\cdot 4R_1\left\|\boldsymbol{z}-\boldsymbol{z}'\right\| + \left\|\boldsymbol{A}\right\|_{\ell^1\to\ell^2}^2\cdot 4R_1\cdot\left\|\boldsymbol{z}-\boldsymbol{z}'\right\|\cdot 2 \\
\le\ & 16\left\|\boldsymbol{A}\right\|_{\ell^1\to\ell^2}^2 R_1\left\|\boldsymbol{z}-\boldsymbol{z}'\right\|,
\end{aligned}
$$

where to obtain the third inequality we used that $\frac{1}{m}\left\|\boldsymbol{A}^*\right\|^2 \le 2$ with probability at least $1-\exp(-c_3 m)$ when $m \ge C_4 n$ for a sufficiently large constant C_4.

Since $R_0 \le 10\left\|\boldsymbol{x}\right\|$ with probability at least $1-\exp(-c_5 m)$ when $m \ge C_6 n$, by definition of R_1, we have $R_1 \le 30(n\log m)^{1/2}\left\|\boldsymbol{x}\right\|$ with high probability. Substituting this estimate into the above bounds yields the claimed results. ■

18.3 Proof of Lemma 14.2

Proof For the upper bound, we have that for all $\boldsymbol{z}\in\mathcal{R}_3'$,

$$
\begin{aligned}
\left\|\boldsymbol{H}(\boldsymbol{z})\right\| \le \left\|\nabla^2 f(\boldsymbol{z})\right\| &\le \left\|\nabla^2 f(\boldsymbol{x}e^{\mathrm{i}\phi(\boldsymbol{z})})\right\| + L_h\left\|\boldsymbol{h}(\boldsymbol{z})\right\| \\
&\le \left\|\nabla^2 f(\boldsymbol{x}e^{\mathrm{i}\phi(\boldsymbol{z})}) - \mathbb{E}\left[\nabla^2 f(\boldsymbol{x}e^{\mathrm{i}\phi(\boldsymbol{z})})\right]\right\| + \left\|\mathbb{E}\left[\nabla^2 f(\boldsymbol{x}e^{\mathrm{i}\phi(\boldsymbol{z})})\right]\right\| + \frac{1}{10}\left\|\boldsymbol{x}\right\|^2 \\
&\le \frac{1}{100}\left\|\boldsymbol{x}\right\|^2 + 4\left\|\boldsymbol{x}\right\|^2 + \frac{1}{10}\left\|\boldsymbol{x}\right\|^2 \le \frac{9}{2}\left\|\boldsymbol{x}\right\|^2,
\end{aligned}
$$

where to obtain the third line we applied Lemma 18.2. To show the lower bound for all $\boldsymbol{z} \in \mathcal{R}_3'$, it is equivalent to show that

$$\frac{1}{2}\begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix} \ge m_H, \quad \forall\ \|\boldsymbol{w}\| = 1 \text{ with } \Im(\boldsymbol{w}^*\boldsymbol{z}) = 0, \text{ and } \forall\ \boldsymbol{z} \in \mathcal{R}_3'.$$

By Lemma 14.1 and Lemma 18.2, with high probability, we have

$$\begin{aligned}\frac{1}{2}\begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix} &\ge \frac{1}{2}\begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{x}e^{\mathrm{i}\phi(\boldsymbol{z})}) \begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix} - L_h \|\boldsymbol{h}(\boldsymbol{z})\| \|\boldsymbol{w}\|^2 \\ &\ge \frac{1}{2}\begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix}^* \mathbb{E}\left[\nabla^2 f(\boldsymbol{x}e^{\mathrm{i}\phi(\boldsymbol{z})})\right] \begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix} - \left(\frac{1}{10} + \frac{1}{100}\right)\|\boldsymbol{x}\|^2 \\ &= \left(1 - \frac{1}{100} - \frac{1}{10}\right)\|\boldsymbol{x}\|^2 + |\boldsymbol{w}^*\boldsymbol{x}|^2 + 2\Re\left((\boldsymbol{w}^*\boldsymbol{x}e^{\mathrm{i}\phi(\boldsymbol{z})})^2\right) \\ &\ge \frac{89}{100}\|\boldsymbol{x}\|^2 + \Re\left((\boldsymbol{w}^*\boldsymbol{x}e^{\mathrm{i}\phi(\boldsymbol{z})})^2\right).\end{aligned}$$

Since $\Im(\boldsymbol{w}^*\boldsymbol{z}) = 0$, we have $\Re\left((\boldsymbol{w}^*\boldsymbol{z})^2\right) = |\boldsymbol{w}^*\boldsymbol{z}|^2$. Thus,

$$\begin{aligned}\Re\left((\boldsymbol{w}^*\boldsymbol{x}e^{\mathrm{i}\phi(\boldsymbol{z})})^2\right) &= \Re\left((\boldsymbol{w}^*\boldsymbol{z} - \boldsymbol{w}^*\boldsymbol{h}(\boldsymbol{z}))^2\right) \\ &= |\boldsymbol{w}^*\boldsymbol{z}|^2 + \Re\left((\boldsymbol{w}^*\boldsymbol{h})^2\right) - 2\Re\left((\boldsymbol{w}^*\boldsymbol{h}(\boldsymbol{z}))(\boldsymbol{w}^*\boldsymbol{z})\right) \\ &\ge |\boldsymbol{w}^*\boldsymbol{z}|^2 - \|\boldsymbol{w}\|^2\|\boldsymbol{h}(\boldsymbol{z})\|^2 - 2\|\boldsymbol{w}\|^2\|\boldsymbol{h}(\boldsymbol{z})\|\|\boldsymbol{z}\| \\ &\ge -\frac{1}{100L_h^2}\|\boldsymbol{x}\|^4 - \frac{2}{10L_h}\|\boldsymbol{x}\|^2\left(\|\boldsymbol{x}\| + \frac{1}{10L_h}\|\boldsymbol{x}\|^2\right) \\ &\ge -\frac{1}{100}\|\boldsymbol{x}\|^2,\end{aligned}$$

where we obtained the last inequality based on the fact that $L_h \doteq 480(n\log m)^{1/2}\|\boldsymbol{A}\|_{\ell^1\to\ell^2}^2\|\boldsymbol{x}\| \ge 150\|\boldsymbol{x}\|$ whenever $\|\boldsymbol{A}\|_{\ell^1\to\ell^2}^2 \ge 1$; this holds with high probability when $m \ge C_1 n$ for large enough constant C_1. Together we obtain

$$\frac{1}{2}\begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix}\boldsymbol{w}\\ \overline{\boldsymbol{w}}\end{bmatrix} \ge \frac{89}{100}\|\boldsymbol{x}\|^2 - \frac{1}{100}\|\boldsymbol{x}\|^2 \ge \frac{22}{25}\|\boldsymbol{x}\|^2,$$

as desired. ■

18.4 Proof of Lemma 14.3

Proof In view of Lemma C.3, we have

$$\begin{aligned} f(\boldsymbol{z}+\boldsymbol{\delta}_\star) &\le \widehat{f}(\boldsymbol{\delta}_\star;\boldsymbol{z}) + \tfrac{1}{3}L_h\Delta^3 \\ &\le \widehat{f}(\boldsymbol{\delta};\boldsymbol{z}) + \tfrac{1}{3}L_h\Delta^3 \\ &\le f(\boldsymbol{z}+\boldsymbol{\delta}) + \tfrac{2}{3}L_h\Delta^3 \\ &\le f(\boldsymbol{z}) - d + \tfrac{2}{3}L_h\Delta^3, \end{aligned}$$

as desired. ■

18.5 Proof of Proposition 14.4

Proof In view of Proposition 13.3, consider direction $\boldsymbol{\delta} \doteq \boldsymbol{x}e^{i\phi(\boldsymbol{z})}/\left\|\boldsymbol{x}\right\|$. Obviously, vectors of the form $t\sigma\boldsymbol{\delta}$ are feasible for (14.1.1) for any $t \in [0,\Delta]$ and $\sigma \doteq -\operatorname{sign}([\boldsymbol{\delta}^*, \overline{\boldsymbol{\delta}}^*]\nabla f(\boldsymbol{z}^{(r)}))$. By Lemma C.2, we obtain

$$\begin{aligned} f(\boldsymbol{z}^{(r)}+t\sigma\boldsymbol{\delta}) &= f(\boldsymbol{z}^{(r)}) + t\sigma\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r)}) + t^2\int_0^1(1-s)\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z}^{(r)}+\sigma st\boldsymbol{\delta})\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix} ds \\ &\le f(\boldsymbol{z}^{(r)}) + \frac{t^2}{2}\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z}^{(r)})\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix} \\ &+ t^2\int_0^1(1-s)\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix}^* \left[\nabla^2 f(\boldsymbol{z}^{(r)}+\sigma st\boldsymbol{\delta}) - \nabla^2 f(\boldsymbol{z}^{(r)})\right]\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix} ds \\ &\le f(\boldsymbol{z}^{(r)}) + \frac{t^2}{2}\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z}^{(r)})\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix} + \frac{L_h}{3}t^3. \end{aligned}$$

Thus, we have

$$f(\boldsymbol{z}^{(r)}+t\sigma\boldsymbol{\delta}) - f(\boldsymbol{z}^{(r)}) \le -\frac{1}{200}t^2\left\|\boldsymbol{x}\right\|^2 + \frac{L_h}{3}t^3.$$

Taking $t=\Delta$ and applying Lemma 14.3, we have

$$f(\boldsymbol{z}^{(r+1)}) - f(\boldsymbol{z}^{(r)}) \le -\frac{1}{200}\Delta^2\left\|\boldsymbol{x}\right\|^2 + \frac{L_h}{3}\Delta^3 + \frac{2}{3}L_h\Delta^3 \le -\frac{1}{200}\Delta^2\left\|\boldsymbol{x}\right\|^2 + L_h\Delta^3 \le -\frac{1}{400}\left\|\boldsymbol{x}\right\|^2\Delta^2,$$

where we obtain the very last inequality using the assumption that $\Delta \leq \|\boldsymbol{x}\|^2/(400L_h)$, completing the proof.

■

18.6 Proof of Proposition 14.5

Proof We take

$$\boldsymbol{\delta} = \begin{cases} -\boldsymbol{z}^{(r)}/\left\|\boldsymbol{z}^{(r)}\right\| & \boldsymbol{z}^{(r)} \in \mathcal{R}_2^{\boldsymbol{z}} \\ -\boldsymbol{h}(\boldsymbol{z}^{(r)})/\left\|\boldsymbol{h}(\boldsymbol{z}^{(r)})\right\| & \boldsymbol{z}^{(r)} \in \mathcal{R}_2^{\boldsymbol{h}} \end{cases}.$$

Obviously vectors of the form $t\boldsymbol{\delta}$ is feasible for (14.1.1) for any $t \in [0, \Delta]$. By Lemma C.2, we have

$$\begin{aligned} f(\boldsymbol{z}^{(r)} + t\boldsymbol{\delta}) &= f(\boldsymbol{z}^{(r)}) + t\int_0^1 \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r)} + st\boldsymbol{\delta})\, ds \\ &= f(\boldsymbol{z}^{(r)}) + t\begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{x}^{(r)}) + t\int_0^1 \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \left[\nabla f(\boldsymbol{z}^{(r)} + st\boldsymbol{\delta}) - \nabla f(\boldsymbol{z}^{(r)})\right]\, ds \\ &\leq f(\boldsymbol{z}^{(r)}) + t\begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r)}) + t^2 L_g. \end{aligned}$$

By Proposition 13.4 and Proposition 13.5, we have

$$f(\boldsymbol{z}^{(r)} + t\boldsymbol{\delta}) - f(\boldsymbol{z}^{(r)}) \leq -\frac{1}{1000} t\|\boldsymbol{x}\|^2 \|\boldsymbol{z}^{(r)}\| + t^2 L_g.$$

Since $\{\boldsymbol{z} : \|\boldsymbol{z}\| \leq \|\boldsymbol{x}\|/2\} \subset \mathcal{R}_1$, $\boldsymbol{z}^{(r)}$ of interest here satisfies $\|\boldsymbol{z}^{(r)}\| \geq \|\boldsymbol{x}\|/2$. Thus,

$$f(\boldsymbol{z}^{(r)} + t\boldsymbol{\delta}) - f(\boldsymbol{z}^{(r)}) \leq -\frac{1}{2000} t\|\boldsymbol{x}\|^3 + t^2 L_g.$$

Combining the above with Lemma 14.3, we obtain

$$f(\boldsymbol{z}^{(r+1)}) - f(\boldsymbol{z}^{(r)}) \leq -\frac{1}{2000}\Delta\|\boldsymbol{x}\|^3 + \Delta^2 L_g + \frac{2}{3}L_h\Delta^3 \leq -\frac{1}{4000}\Delta\|\boldsymbol{x}\|^3,$$

provided

$$\Delta \leq \min\left\{\frac{\|\boldsymbol{x}\|^3}{8000L_g}, \sqrt{\frac{3\|\boldsymbol{x}\|^3}{16000L_h}}\right\},$$

as desired. ■

18.7 Proof of Proposition 14.6

Proof By Proposition 13.6 and the integral form of Taylor's theorem in Lemma C.2, we have that for any $\boldsymbol{g}$ satisfying $\Im(\boldsymbol{g}^*\boldsymbol{x}) = 0$ and $\|\boldsymbol{g}\| = 1$ and any $t \in [0, \|\boldsymbol{x}\| / \sqrt{7}]$,

$$\begin{aligned} f(\boldsymbol{x} + t\boldsymbol{g}) &= f(\boldsymbol{x}) + t \begin{bmatrix} \boldsymbol{g} \\ \overline{\boldsymbol{g}} \end{bmatrix}^* \nabla f(\boldsymbol{x}) + t^2 \int_0^1 (1-s) \begin{bmatrix} \boldsymbol{g} \\ \overline{\boldsymbol{g}} \end{bmatrix}^* \nabla^2 f(\boldsymbol{x} + st\boldsymbol{g}) \begin{bmatrix} \boldsymbol{g} \\ \overline{\boldsymbol{g}} \end{bmatrix} \, ds \\ &\geq f(\boldsymbol{x}) + t \begin{bmatrix} \boldsymbol{g} \\ \overline{\boldsymbol{g}} \end{bmatrix}^* \nabla f(\boldsymbol{x}) + \frac{1}{8} \|\boldsymbol{x}\|^2 t^2. \end{aligned}$$

Similarly, we have

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x} + t\boldsymbol{g}) - t \begin{bmatrix} \boldsymbol{g} \\ \overline{\boldsymbol{g}} \end{bmatrix}^* \nabla f(\boldsymbol{x} + t\boldsymbol{g}) + \frac{1}{8} \|\boldsymbol{x}\|^2 t^2.$$

Combining the above two inequalities, we obtain

$$t \begin{bmatrix} \boldsymbol{g} \\ \overline{\boldsymbol{g}} \end{bmatrix}^* (\nabla f(\boldsymbol{x} + t\boldsymbol{g}) - \nabla f(\boldsymbol{x})) \geq \frac{1}{4} \|\boldsymbol{x}\|^2 t^2 \Longrightarrow \begin{bmatrix} \boldsymbol{g} \\ \overline{\boldsymbol{g}} \end{bmatrix}^* \nabla f(\boldsymbol{x} + t\boldsymbol{g}) \geq \frac{1}{4} \|\boldsymbol{x}\|^2 t \geq \frac{1}{40 L_h} \|\boldsymbol{x}\|^4,$$

where to obtain the very last bound we have used the fact $\min_{\boldsymbol{z} \in \mathcal{R}_3 \setminus \mathcal{R}_3'} \|\boldsymbol{h}(\boldsymbol{z})\| \geq \|\boldsymbol{x}\|^2 / (10 L_h)$ due to (14.2.3). This implies that for all $\boldsymbol{z} \in \mathcal{R}_3 \setminus \mathcal{R}_3'$,

$$\begin{bmatrix} \boldsymbol{h}(\boldsymbol{z}) \\ \overline{\boldsymbol{h}(\boldsymbol{z})} \end{bmatrix}^* \nabla f(\boldsymbol{z}) \geq \frac{1}{40 L_h} \|\boldsymbol{x}\|^4. \tag{18.7.1}$$

The rest arguments are very similar to that of Proposition 14.5. Take $\boldsymbol{\delta} = -\boldsymbol{h}(\boldsymbol{z}^{(r)}) / \left\|\boldsymbol{h}(\boldsymbol{z}^{(r)})\right\|$ and it can checked vectors of the form $t\boldsymbol{\delta}$ for $t \in [0, \Delta]$ are feasible for (14.1.1). By Lemma C.2, we have

$$\begin{aligned} f(\boldsymbol{z}^{(r)} + t\boldsymbol{\delta}) &= f(\boldsymbol{z}^{(r)}) + t \int_0^1 \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r)} + st\boldsymbol{\delta}) \, ds \\ &= f(\boldsymbol{z}^{(r)}) + t \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{x}^{(r)}) + t \int_0^1 \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \left[\nabla f(\boldsymbol{z}^{(r)} + st\boldsymbol{\delta}) - \nabla f(\boldsymbol{z}^{(r)}) \right] \, ds \\ &\leq f(\boldsymbol{z}^{(r)}) + t \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r)}) + t^2 L_g \end{aligned}$$

$$\leq f(\boldsymbol{z}^{(r)}) - \frac{1}{40L_h} t \left\| \boldsymbol{x} \right\|^4 + t^2 L_g,$$

where to obtain the last line we have used (18.7.1). Combining the above with Lemma 14.3, we obtain

$$f(\boldsymbol{z}^{(r+1)}) - f(\boldsymbol{z}^{(r)}) \leq -\frac{1}{40L_h} \Delta \left\| \boldsymbol{x} \right\|^4 + \Delta^2 L_g + \frac{2}{3} L_h \Delta^3 \leq -\frac{1}{80L_h} \Delta \left\| \boldsymbol{x} \right\|^4,$$

provided

$$\Delta \leq \min \left\{ \frac{\left\| \boldsymbol{x} \right\|^4}{160 L_h L_g}, \sqrt{\frac{3}{320}} \frac{\left\| \boldsymbol{x} \right\|^2}{L_h} \right\},$$

as desired. ■

18.8 Proof of Proposition 14.7

Proof If we identify $\mathbb{C}^n$ with $\mathbb{R}^{2n}$, it can be easily verified that the orthoprojectors of a vector $\boldsymbol{w}$ onto $\boldsymbol{z}$ and its orthogonal complement are

$$\mathcal{P}_{\boldsymbol{z}}(\boldsymbol{w}) = \frac{\Re(\boldsymbol{z}^* \boldsymbol{w}) \boldsymbol{z}}{\left\| \boldsymbol{z} \right\|^2}, \quad \text{and} \quad \mathcal{P}_{\boldsymbol{z}^\perp}(\boldsymbol{w}) = \boldsymbol{w} - \frac{\Re(\boldsymbol{z}^* \boldsymbol{w}) \boldsymbol{z}}{\left\| \boldsymbol{z} \right\|^2}.$$

Now at any point $\boldsymbol{z}^{(r)} \in \mathcal{R}_3'$, consider a feasible direction of the form $\boldsymbol{\delta} \doteq -t \mathcal{P}_{(\mathrm{i}\boldsymbol{z}^{(r)})^\perp} \nabla_{\boldsymbol{z}^{(r)}} f(\boldsymbol{z}^{(r)})$ $(0 \leq t \leq \Delta / \| \mathcal{P}_{(\mathrm{i}\boldsymbol{z}^{(r)})^\perp} \nabla_{\boldsymbol{z}^{(r)}} f(\boldsymbol{z}^{(r)}) \|)$ to the trust-region subproblem (14.1.1). The local quadratic approximation obeys

$$\begin{aligned}
\widehat{f}(\boldsymbol{\delta}; \boldsymbol{z}^{(r)}) &= f(\boldsymbol{z}^{(r)}) + \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r)}) + \frac{1}{2} \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}^{(r)}) \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix} \\
&\leq f(\boldsymbol{z}^{(r)}) - 2t \left\| \mathcal{P}_{(\mathrm{i}\boldsymbol{z}^{(r)})^\perp} \nabla_{\boldsymbol{z}^{(r)}} f(\boldsymbol{z}^{(r)}) \right\|^2 + t^2 M_H \left\| \mathcal{P}_{(\mathrm{i}\boldsymbol{z}^{(r)})^\perp} \nabla_{\boldsymbol{z}^{(r)}} f(\boldsymbol{z}^{(r)}) \right\|^2 \\
&= f(\boldsymbol{z}^{(r)}) - 2t \left(1 - \frac{M_H}{2} t \right) \left\| \mathcal{P}_{(\mathrm{i}\boldsymbol{z}^{(r)})^\perp} \nabla_{\boldsymbol{z}^{(r)}} f(\boldsymbol{z}^{(r)}) \right\|^2,
\end{aligned}$$

where M_H is as defined in Lemma 14.2. Taking $t = \min\{M_H^{-1}, \Delta / \| \mathcal{P}_{(\mathrm{i}\boldsymbol{z}^{(r)})^\perp} \nabla_{\boldsymbol{z}^{(r)}} f(\boldsymbol{z}^{(r)}) \| \}$, we have

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{z}^{(r)}) - f(\boldsymbol{z}^{(r)}) \leq -\min\{M_H^{-1}, \Delta / \| \mathcal{P}_{(\mathrm{i}\boldsymbol{z}^{(r)})^\perp} \nabla_{\boldsymbol{z}^{(r)}} f(\boldsymbol{z}^{(r)}) \| \} \left\| \mathcal{P}_{(\mathrm{i}\boldsymbol{z}^{(r)})^\perp} \nabla_{\boldsymbol{z}^{(r)}} f(\boldsymbol{z}^{(r)}) \right\|^2.$$

Let $\boldsymbol{U}$ be an orthogonal (in geometric sense) basis for the space $\{ \boldsymbol{w} : \Im(\boldsymbol{w}^* \boldsymbol{z}^{(r)}) = 0 \}$. In view of the transformed gradient and Hessian in (14.1.3), it is easy to see

$$\left\| \mathcal{P}_{(\mathrm{i}\boldsymbol{z}^{(r)})^\perp} \nabla_{\boldsymbol{z}^{(r)}} f(\boldsymbol{z}^{(r)}) \right\| = \frac{1}{\sqrt{2}} \left\| \boldsymbol{g}(\boldsymbol{z}^{(r)}) \right\|,$$

where $\boldsymbol{g}(\boldsymbol{z}^{(r)})$ is the transformed gradient. To lower bound $\left\|\mathcal{P}_{(\mathrm{i}\boldsymbol{z}^{(r)})^\perp}\nabla_{\boldsymbol{z}^{(r)}} f(\boldsymbol{z}^{(r)})\right\|$, recall the step is constrained, we have

$$\Delta \le \left\|\boldsymbol{H}^{-1}(\boldsymbol{z}^{(r)})\boldsymbol{g}(\boldsymbol{z}^{(r)})\right\| \le \left\|\boldsymbol{H}^{-1}(\boldsymbol{z}^{(r)})\right\| \left\|\boldsymbol{g}(\boldsymbol{z}^{(r)})\right\| \le \frac{1}{\lambda_{\min}(\boldsymbol{H}(\boldsymbol{z}^{(r)}))}\left\|\boldsymbol{g}(\boldsymbol{z}^{(r)})\right\|.$$

By Lemma 14.2, $\lambda_{\min}(\boldsymbol{H}(\boldsymbol{z}^{(r)})) \ge m_H$. Thus,

$$\left\|\boldsymbol{g}(\boldsymbol{z}^{(r)})\right\| \ge m_H\Delta.$$

Hence we have

$$\widehat{f}(\boldsymbol{\delta};\boldsymbol{z}^{(r)}) - f(\boldsymbol{z}^{(r)}) \le -\min\left\{\frac{m_H^2\Delta^2}{2M_H}, \frac{\Delta^2 m_H}{\sqrt{2}}\right\} \le -\frac{m_H^2\Delta^2}{2M_H},$$

where the last simplification is due to that $M_H \ge m_H$. By Lemma C.3, we have

$$f(\boldsymbol{z}^{(r)}+\boldsymbol{\delta}) - f(\boldsymbol{z}^{(r)}) \le -\frac{m_H^2\Delta^2}{2M_H} + \frac{L_h}{3}\Delta^3.$$

Therefore, for $\boldsymbol{z}^{(r+1)} = \boldsymbol{z}^{(r)} + \boldsymbol{\delta}_\star$, Lemma 14.3 implies that

$$f(\boldsymbol{z}^{(r+1)}) - f(\boldsymbol{z}^{(r)}) \le -\frac{m_H^2\Delta^2}{2M_H} + L_h\Delta^3.$$

The claimed result follows provided $\Delta \le m_H^2/(4M_H L_h)$, completing the proof. ■

18.9 Proof of Proposition 14.8

Before proceeding, we note one important fact that is useful below. For any $\boldsymbol{z}$, we have

$$\mathcal{P}_{\mathrm{i}\boldsymbol{z}}\nabla_{\boldsymbol{z}} f(\boldsymbol{z}) = \frac{\Re((\mathrm{i}\boldsymbol{z})^*\nabla_{\boldsymbol{z}} f(\boldsymbol{z}))}{\|\boldsymbol{z}\|^2}\mathrm{i}\boldsymbol{z} = \mathbf{0}.$$

Thus, if $\boldsymbol{U}(\boldsymbol{z})$ is an (geometrically) orthonormal basis constructed for the space $\{\boldsymbol{w} : \Im(\boldsymbol{w}^*\boldsymbol{z}) = 0\}$ (as defined around (14.1.3)), it is easy to verify that

$$\begin{bmatrix}\boldsymbol{U}\\ \overline{\boldsymbol{U}}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}\\ \overline{\boldsymbol{U}}\end{bmatrix}^* \nabla f(\boldsymbol{z}) = 2\nabla f(\boldsymbol{z}). \tag{18.9.1}$$

We next prove Proposition 14.8.

Proof Throughout the proof, we write $\boldsymbol{g}^{(r)}$, $\boldsymbol{H}^{(r)}$ and $\boldsymbol{U}^{(r)}$ short for $\boldsymbol{g}(\boldsymbol{z}^{(r)})$, $\boldsymbol{H}(\boldsymbol{z}^{(r)})$ and $\boldsymbol{U}(\boldsymbol{z}^{(r)})$, respectively. Given an orthonormal basis $\boldsymbol{U}^{(r)}$ for $\left\{\boldsymbol{w} : \Im(\boldsymbol{w}^*\boldsymbol{z}^{(r)}) = 0\right\}$, the unconstrained optimality condition of the trust region method implies that

$$\boldsymbol{H}^{(r)}\boldsymbol{\xi}_\star + \boldsymbol{g}^{(r)} = \boldsymbol{0} \Longleftrightarrow \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}^{(r)}) \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \boldsymbol{\xi}_\star + \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r)}) = \boldsymbol{0}.$$

Thus, we have

$$\begin{aligned}
&\|\nabla f(\boldsymbol{z}^{(r+1)})\| \\
=& \frac{1}{2}\left\| \begin{bmatrix} \boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}} \end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r+1)}) \right\| \\
=& \frac{1}{2}\left\| \begin{bmatrix} \boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}} \end{bmatrix}^* \nabla f(\boldsymbol{z}^{(r+1)}) - \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix}^* \left(\nabla^2 f(\boldsymbol{z}^{(r)}) \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \boldsymbol{\xi}_\star + \nabla f(\boldsymbol{z}^{(r)}) \right) \right\| \\
\le& \frac{1}{2}\left\| \begin{bmatrix} \boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}} \end{bmatrix}^* \left[\nabla f(\boldsymbol{z}^{(r+1)}) - \nabla f(\boldsymbol{z}^{(r)}) - \nabla^2 f(\boldsymbol{z}^{(r)}) \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \boldsymbol{\xi}_\star \right] \right\| \\
&\quad + \frac{1}{2}\left\| \left(\begin{bmatrix} \boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}} \end{bmatrix}^* - \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix}^* \right) \left(\nabla^2 f(\boldsymbol{z}^{(r)}) \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \boldsymbol{\xi}_\star + \nabla f(\boldsymbol{z}^{(r)}) \right) \right\| \\
\le& \left\| \nabla f(\boldsymbol{z}^{(r+1)}) - \nabla f(\boldsymbol{z}^{(r)}) - \nabla^2 f(\boldsymbol{z}^{(r)}) \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \boldsymbol{\xi}_\star \right\| \\
&\quad + \frac{1}{2}\left\| \begin{bmatrix} \boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}} \end{bmatrix}^* - \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix}^* \right\| \left\| \nabla^2 f(\boldsymbol{z}^{(r)}) \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \boldsymbol{\xi}_\star + \nabla f(\boldsymbol{z}^{(r)}) \right\|.
\end{aligned}$$

By Taylor's theorem and Lipschitz property in Lemma 14.1, we have

$$\begin{aligned}
&\left\| \nabla f(\boldsymbol{z}^{(r+1)}) - \nabla f(\boldsymbol{z}^{(r)}) - \nabla^2 f(\boldsymbol{z}^{(r)}) \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \boldsymbol{\xi}_\star \right\| \\
=& \left\| \int_0^1 \left[\nabla^2 f(\boldsymbol{z}^{(r)} + t \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \boldsymbol{\xi}_\star) - \nabla^2 f(\boldsymbol{z}^{(r)}) \right] \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \boldsymbol{\xi}_\star \, dt \right\| \\
\le& \|\boldsymbol{\xi}_\star\| \int_0^1 \left\| \nabla^2 f(\boldsymbol{z}^{(r)} + t \begin{bmatrix} \boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}} \end{bmatrix} \boldsymbol{\xi}_\star) - \nabla^2 f(\boldsymbol{z}^{(r)}) \right\| \, dt \\
\le& \frac{1}{2} L_h \left\| \boldsymbol{\xi}_\star \right\|^2 .
\end{aligned} \tag{18.9.2}$$

Moreover,

$$\begin{aligned}\left\|\nabla f(\boldsymbol{z}^{(r)})\right\| &= \frac{1}{\sqrt{2}}\left\|\begin{bmatrix}\boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}}\end{bmatrix}^{*}\nabla f(\boldsymbol{z}^{(r)})\right\| \\ &= \frac{1}{\sqrt{2}}\left\|-\begin{bmatrix}\boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}}\end{bmatrix}^{*}\nabla^2 f(\boldsymbol{z}^{(r)})\begin{bmatrix}\boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}}\end{bmatrix}\boldsymbol{\xi}_\star\right\| \le \sqrt{2}\left\|\nabla^2 f(\boldsymbol{z}^{(r)})\right\|\left\|\boldsymbol{\xi}_\star\right\|,\end{aligned}$$

where to obtain the second equality we have used the optimality condition discussed at start of the proof. Thus, using the result above, we obtain

$$\left\|\nabla^2 f(\boldsymbol{z}^{(r)})\begin{bmatrix}\boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}}\end{bmatrix}\boldsymbol{\xi}_\star + \nabla f(\boldsymbol{z}^{(r)})\right\| \le 2\sqrt{2}\left\|\nabla^2 f(\boldsymbol{z}^{(r)})\right\|\left\|\boldsymbol{\xi}_\star\right\|. \tag{18.9.3}$$

On the other hand,

$$\begin{aligned}&\left\|\begin{bmatrix}\boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}^{(r+1)} \\ \overline{\boldsymbol{U}^{(r+1)}}\end{bmatrix}^{*} - \begin{bmatrix}\boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}^{(r)} \\ \overline{\boldsymbol{U}^{(r)}}\end{bmatrix}^{*}\right\| \\ \le\ &\left\|\boldsymbol{U}^{(r+1)}(\boldsymbol{U}^{(r+1)})^{*} - \boldsymbol{U}^{(r)}(\boldsymbol{U}^{(r)})^{*}\right\| + \left\|\boldsymbol{U}^{(r+1)}(\boldsymbol{U}^{(r+1)})^{\top} - \boldsymbol{U}^{(r)}(\boldsymbol{U}^{(r)})^{\top}\right\|.\end{aligned}$$

Write $\boldsymbol{U}^{(r+1)} = \boldsymbol{U}_{\Re}^{(r+1)} + \mathrm{i}\boldsymbol{U}_{\Im}^{(r+1)}$, where $\boldsymbol{U}_{\Re}^{(r+1)}$ and $\boldsymbol{U}_{\Im}^{(r+1)}$ collect respectively entrywise real and imaginary parts of $\boldsymbol{U}^{(r+1)}$. It is not difficult to verify that $\boldsymbol{V}^{(r+1)} \doteq [\boldsymbol{U}_{\Re}^{(r+1)}; \boldsymbol{U}_{\Im}^{(r+1)}] \in \mathbb{R}^{2n\times(2n-1)}$ is an orthonormal matrix. We also define $\boldsymbol{V}^{(r)}$ accordingly. Thus,

$$\begin{aligned}\left\|\boldsymbol{U}^{(r+1)}(\boldsymbol{U}^{(r+1)})^{*} - \boldsymbol{U}^{(r)}(\boldsymbol{U}^{(r)})^{*}\right\| &= \left\|[\boldsymbol{I}, \mathrm{i}\boldsymbol{I}]\left(\boldsymbol{V}^{(r+1)}(\boldsymbol{V}^{(r+1)})^{\top} - \boldsymbol{V}^{(r)}(\boldsymbol{V}^{(r)})^{\top}\right)[\boldsymbol{I}, -\mathrm{i}\boldsymbol{I}]^{\top}\right\| \\ &\le 2\left\|\boldsymbol{V}^{(r+1)}(\boldsymbol{V}^{(r+1)})^{\top} - \boldsymbol{V}^{(r)}(\boldsymbol{V}^{(r)})^{\top}\right\| \\ &\le 2\sqrt{2}\left\|\boldsymbol{V}^{(r+1)}(\boldsymbol{V}^{(r+1)})^{\top} - \boldsymbol{V}^{(r)}(\boldsymbol{V}^{(r)})^{\top}\right\|_{R},\end{aligned}$$

where from the second to the third line we translate the complex operator norm to the real operator norm. Similarly, we also get

$$\left\|\boldsymbol{U}^{(r+1)}(\boldsymbol{U}^{(r+1)})^{\top} - \boldsymbol{U}^{(r)}(\boldsymbol{U}^{(r)})^{\top}\right\| \le 2\sqrt{2}\left\|\boldsymbol{V}^{(r+1)}(\boldsymbol{V}^{(r+1)})^{\top} - \boldsymbol{V}^{(r)}(\boldsymbol{V}^{(r)})^{\top}\right\|_{R}.$$

Since $\mathrm{i}\boldsymbol{z}^{(r)}$ is the normal vector of the space generated by $\boldsymbol{U}^{(r)}$, $[-\boldsymbol{z}_{\Im}^{(r)}; \boldsymbol{z}_{\Re}^{(r)}]$ is the corresponding normal vector of $\boldsymbol{V}^{(r)}$. By Lemma B.13, the largest principal angle θ_1 between the subspaces designated by $\boldsymbol{V}^{(r+1)}$ and $\boldsymbol{V}^{(r)}$ are the angle between their normal vectors $\boldsymbol{a} \doteq [-\boldsymbol{z}_{\Im}^{(r)}; \boldsymbol{z}_{\Re}^{(r)}]$ and $\boldsymbol{b} \doteq [-\boldsymbol{z}_{\Im}^{(r+1)}; \boldsymbol{z}_{\Re}^{(r+1)}]$. Here we have decomposed $\boldsymbol{z}^{(r+1)}$ and $\boldsymbol{z}^{(r)}$ into real and imaginary parts. Similarly we define $\boldsymbol{c} \doteq [-(\boldsymbol{\delta}_\star)_\Im; (\boldsymbol{\delta}_\star)_\Re]$. By the law of cosines,

$$\cos\theta_1 = \frac{\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{c}\|^2}{2\|\boldsymbol{a}\|\|\boldsymbol{b}\|} \geq 1 - \frac{\|\boldsymbol{c}\|^2}{2\|\boldsymbol{a}\|\|\boldsymbol{b}\|} = 1 - \frac{\|\boldsymbol{\xi}_\star\|^2}{2\left\|\boldsymbol{z}^{(r)}\right\|\left\|\boldsymbol{z}^{(r+1)}\right\|}.$$

Since $\left\|\boldsymbol{z}^{(r)}\right\| \geq \min_{\boldsymbol{z}\in\mathcal{R}_3}\|\boldsymbol{z}\| \geq (1-1/\sqrt{7})\|\boldsymbol{x}\| \geq 3\|\boldsymbol{x}\|/5$, and $\left\|\boldsymbol{z}^{(r+1)}\right\| \geq \left\|\boldsymbol{z}^{(r)}\right\| - \Delta \geq \|\boldsymbol{x}\|/2$ provided

$$\Delta \leq \|\boldsymbol{x}\|/10,$$

we obtain that

$$\cos\theta_1 \geq 1 - \frac{5}{3\|\boldsymbol{x}\|^2}\|\boldsymbol{\xi}_\star\|^2.$$

Thus, by Lemma B.13 again,

$$\begin{aligned}\left\|\boldsymbol{V}^{(r+1)}(\boldsymbol{V}^{(r+1)})^\top - \boldsymbol{V}^{(r)}(\boldsymbol{V}^{(r)})^\top\right\|_R &= \sqrt{1-\cos^2\theta_1} \\ &\leq \sqrt{\frac{10}{3\|\boldsymbol{x}\|^2}\|\boldsymbol{\delta}_\star\|^2 + \frac{25}{9\|\boldsymbol{x}\|^4}\|\boldsymbol{\xi}_\star\|^4} \leq \frac{2}{\|\boldsymbol{x}\|}\|\boldsymbol{\xi}_\star\|, \quad (18.9.4)\end{aligned}$$

where we used the assumption $\Delta \leq \|\boldsymbol{x}\|/10$ again to obtain the last inequality.

Collecting the above results, we obtain

$$\left\|\nabla f(\boldsymbol{z}^{(r+1)})\right\| \leq \left(\frac{1}{2}L_h + \frac{16}{\|\boldsymbol{x}\|}M_H\right)\|\boldsymbol{\xi}_\star\|^2. \quad (18.9.5)$$

Invoking the optimality condition again, we obtain

$$\|\boldsymbol{\xi}_\star\|^2 = \left\|(\boldsymbol{H}^{(r)})^{-1}\boldsymbol{g}^{(r)}\right\|^2 \leq \frac{1}{m_H^2}\left\|\boldsymbol{g}^{(r)}\right\|^2 = \frac{2}{m_H^2}\left\|\nabla f(\boldsymbol{z}^{(r)})\right\|^2. \quad (18.9.6)$$

Here $(\boldsymbol{H}^{(r)})^{-1}$ is well-defined because Lemma 14.2 shows that $\left\|\boldsymbol{H}^{(r)}\right\| \geq m_H$ for all $\boldsymbol{z}^{(r)} \in \mathcal{R}_3'$. Combining the last two estimates, we complete the proof. ■

18.10 Proof of Proposition 14.9

Proof Throughout the proof, we write $\boldsymbol{g}^{(r)}$, $\boldsymbol{H}^{(r)}$ and $\boldsymbol{U}^{(r)}$ short for $\boldsymbol{g}(\boldsymbol{z}^{(r)})$, $\boldsymbol{H}(\boldsymbol{z}^{(r)})$ and $\boldsymbol{U}(\boldsymbol{z}^{(r)})$, respectively.

We first show $\boldsymbol{z}^{(r+1)}$ stays in $\mathcal{R}_3'$. From proof of Proposition 14.6, we know that for all $\boldsymbol{z} \in \mathcal{R}_3$, the following estimate holds:

$$\|\nabla f(\boldsymbol{z})\| \geq \frac{1}{4\sqrt{2}} \|\boldsymbol{x}\|^2 \|\boldsymbol{h}(\boldsymbol{z})\| .$$

From Proposition 14.8, we know that

$$\|\nabla f(\boldsymbol{z}^{(r+1)})\| \leq \frac{1}{m_H^2} \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H \right) \|\nabla f(\boldsymbol{z}^{(r)})\|^2$$

provided $\Delta \leq \|\boldsymbol{x}\| /10$. Moreover,

$$\|\nabla f(\boldsymbol{z}^{(r)})\|^2 = \frac{1}{2} \left\|\boldsymbol{g}^{(r)}\right\|^2 \leq M_H^2 \left\|(\boldsymbol{H}^{(r)})^{-1}\boldsymbol{g}^{(r)}\right\|^2 \leq M_H^2 \Delta^2,$$

where the last inequality followed because step r is unconstrained. Combining the above estimates, we obtain that

$$\|\nabla f(\boldsymbol{z}^{(r+1)})\| \leq \frac{1}{m_H^2} \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H \right) M_H^2 \Delta^2.$$

Thus,

$$\left\|\boldsymbol{h}(\boldsymbol{z}^{(r+1)})\right\| \leq \frac{4\sqrt{2}}{\|\boldsymbol{x}\|^2} \|\nabla f(\boldsymbol{z}^{(r+1)})\| \leq \frac{4\sqrt{2}}{m_H^2 \|\boldsymbol{x}\|^2} \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H \right) M_H^2 \Delta^2.$$

So, provided

$$\frac{4\sqrt{2}}{m_H^2 \|\boldsymbol{x}\|^2} \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H \right) M_H^2 \Delta^2 \leq \frac{1}{10 L_h} \|\boldsymbol{x}\|^2 ,$$

$\boldsymbol{z}^{(r+1)}$ stays in $\mathcal{R}_3'$.

Next we show the next step will also be an unconstrained step when Δ is sufficiently small. We have

$$\begin{aligned}
&\|(\boldsymbol{H}^{(r+1)})^{-1}\boldsymbol{g}^{(r+1)}\| \\
\leq\ & \frac{1}{m_H} \|\boldsymbol{g}^{(r+1)}\| = \frac{\sqrt{2}}{m_H} \|\nabla f(\boldsymbol{z}^{(r+1)})\| \\
\leq\ & \frac{\sqrt{2}}{m_H^3} \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H) \right) \|\nabla f(\boldsymbol{z}^{(r)})\|^2 = \frac{1}{\sqrt{2} m_H^3} \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H \right) \|\boldsymbol{g}^{(r)}\|^2 \\
\leq\ & \frac{M_H^2}{\sqrt{2} m_H^3} \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H \right) \|(\boldsymbol{H}^{(r)})^{-1}\boldsymbol{g}^{(r)}\|^2 \leq \frac{M_H^2}{\sqrt{2} m_H^3} \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H \right) \Delta^2,
\end{aligned}$$

where we again applied results of Proposition 14.8 to obtain the third line, and applied the optimality condition to obtain the fourth line. Thus, whenever

$$\frac{M_H^2}{\sqrt{2} m_H^3} \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H \right) \Delta < 1,$$

the transformed trust-region subproblem has its minimizer $\boldsymbol{\xi}^{(r+1)}$ with $\|\boldsymbol{\xi}^{(r+1)}\| < \Delta$. This implies the minimizer $\boldsymbol{\delta}^{(r+1)}$ to the original trust-region subproblem satisfies $\boldsymbol{\delta}^{(r+1)} < \Delta$, as $\|\boldsymbol{\delta}^{r+1}\| = \|\boldsymbol{\xi}^{(r+1)}\|$. Thus, under the above condition the $(r+1)$-th step is also unconstrained.

Repeating the above arguments for all future steps implies that all future steps will be constrained within $\mathcal{R}_3'$.

We next provide an explicit estimate for the rate of convergence in terms of distance of the iterate to the target set X. Again by Proposition 14.8,

$$\begin{aligned}
\|\nabla f(\boldsymbol{z}^{(r+r')})\| &\le m_H^2 \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H\right)^{-1} \left(\frac{1}{m_H^2}\left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H\right)\left\|\nabla f(\boldsymbol{z}^{(r)})\right\|\right)^{2^{r'}} \\
&\le m_H^2 \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H\right)^{-1} \left(\frac{1}{\sqrt{2} m_H^2}\left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H\right)\left\|\boldsymbol{g}^{(r)}\right\|\right)^{2^{r'}} \\
&\le m_H^2 \left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H\right)^{-1} \left(\frac{M_H}{\sqrt{2} m_H^2}\left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H\right)\Delta\right)^{2^{r'}}.
\end{aligned}$$

Thus, provided

$$\frac{M_H}{\sqrt{2} m_H^2}\left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H\right)\Delta \le \frac{1}{2},$$

we have

$$\left\|\boldsymbol{h}(\boldsymbol{z}^{(r+r')})\right\| \le \frac{4\sqrt{2}}{\|\boldsymbol{x}\|^2}\|\nabla f(\boldsymbol{z}^{(r+r')})\| \le \frac{4\sqrt{2} m_H^2}{\|\boldsymbol{x}\|^2}\left(L_h + \frac{32}{\|\boldsymbol{x}\|} M_H\right)^{-1} 2^{-2^{r'}},$$

as claimed. ■

Part IV

Discussion

Chapter 19

Other Problems in the $\mathcal{X}$ Family

> All truths are easy to understand once they are discovered; the point is to discover them.
>
> Galileo Galilei

In this chapter, we describe two more practical examples that also lie in the $\mathcal{X}$ family. They are worked out by other authors, and both arise from signal processing and machine learning applications.

19.1 Orthogonal tensor decomposition

Tensors can be thought of as high-order (i.e., multi-dimensional) arrays, of which matrices are 2-nd order examples. Here we shall focus on 4-th order tensors. If $\mathcal{T} \in \mathbb{R}^{n^4}$ is a 4-th order tensor, we use $\mathcal{T}_{i,j,k,\ell}$ to denote its (i,j,k,ℓ)-th entry. Tensors can be constructed from generalized outer products. Let $\otimes$ denote the normal outer product, such at $[\boldsymbol{u} \otimes v]_{i,j} = u_i v_j$. We define $\boldsymbol{u}^{n^{\otimes 4}}$ as

$$\left[\boldsymbol{u}^{n^{\otimes 4}}\right]_{i,j,k,\ell} \doteq u_i u_j u_k u_\ell.$$

Here we are especially interested in 4-th order tensors $\mathcal{T} \in \mathbb{R}^{n^4}$ of the form

$$\mathcal{T} = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i^{\otimes 4}, \quad \text{with } \boldsymbol{u}_i^\top \boldsymbol{u}_j = \delta_{ij}, \forall\, i,j \in [r], \tag{19.1.1}$$

which are called *orthogonally decomposable tensors*. The reason for this name is that not all tensors can be written in this form, even if they are symmetric: $\mathcal{T}_{i,j,k,\ell} = \mathcal{T}_{\pi(i,j,k,\ell)}$ for any permutation $\boldsymbol{\pi}$ and any set $i,j,k,\ell \in [n]$ [KB09, HL13]. The proper inclusion of orthogonally decomposable tensors in symmetric tensors stands in contrast to the matrix case: symmetric matrices are all orthogonally decomposable by spectral theory. The reason for focusing attention on this restricted class is that a number of problems involving learning latent variable models in machine learning can be cast as decomposition of such tensors:

Given $\mathcal{T}$ in form (19.1.1), find the numbers λ_i's and components $\boldsymbol{u}_i$'s (up to sign and permutation).

Examples include learning mixture of Gaussians, independent component analysis (ICA), hidden Markov models, and so on; see [AGH$^+$14] and the references therein.

Here we further restrict ourselves to the case $\lambda_i = 1$ for all i and $r = n$. Before we present two formulations for the decomposition problem, we need to understand how tensors acting on vectors. For $\mathcal{T}$ in form (19.1.1) with $\lambda_i = 1$ for all i, its action on vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^n$ is defined as

$$\mathcal{T}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}) \doteq \sum_{i \in [n]} (\boldsymbol{u}_i^\top \boldsymbol{x})(\boldsymbol{u}_i^\top \boldsymbol{y})(\boldsymbol{u}_i^\top \boldsymbol{z})(\boldsymbol{u}_i^\top \boldsymbol{w}). \tag{19.1.2}$$

Recall the familiar Rayleigh quotient formulation for eigen-decomposition of symmetric matrices. A natural analog for the orthogonal tensor decomposition problem is

$$\text{minimize}\, f(\boldsymbol{v}) \doteq -\mathcal{T}(\boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}) = -\sum_{i=1}^{n} (\boldsymbol{u}_i^\top \boldsymbol{v})^4 \quad \text{subject to} \quad \|\boldsymbol{v}\|_2 = 1. \tag{19.1.3}$$

[GHJY15] showed (Section C.1.) [1] that $f(\boldsymbol{v})$ in the above formulation has $\pm\boldsymbol{u}_i$'s as its only local and also global minimizers, and the function f is $(7/n, 1/\mathrm{poly}(n), 3, 1/\mathrm{poly}(n))$-ridable over $\mathbb{S}^{n-1}$. Once one of the component is obtained, one can apply deflation to obtain the others, similar to the matrix case.

The deflation trick is however very delicate and noise sensitive to deploy for tensors, as compared to the matrix case. Empirically, obtaining all the components in one shot is preferred. This motivates the second formulation, which is less intuitive to grasp:

$$\text{minimize}\, g(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \doteq \sum_{i \neq j} \mathcal{T}(\boldsymbol{v}_i, \boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{v}_j) = \sum_{i \neq j} \sum_{k \in [n]} (\boldsymbol{u}_k^\top \boldsymbol{v}_i)^2 (\boldsymbol{u}_k^\top \boldsymbol{v}_j)^2,$$

$$\text{subject to } \|\boldsymbol{v}_k\| = 1 \;\forall k \in [n].$$

It is obvious that $g(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) \geq 0$ and the optimal value is 0. One might expect that an explicit orthogonality constraint be enforced to ensure $\boldsymbol{v}_i$'s be mutually orthogonal, which is mysteriously missing here. This can be intuitively understood from the "contrast" terms: $\sum_{k \in [n]} (\boldsymbol{u}_k^\top \boldsymbol{v}_i)^2 (\boldsymbol{u}_k^\top \boldsymbol{v}_j)^2$, which is 0 only when $\boldsymbol{v}_i$ is

[1] [GHJY15] has not used the manifold language as we use here, but resorted to Lagrange multiplier and optimality of the Lagrangian function. For the two decomposition formulations we discussed here, one can verify that the gradient and Hessian they defined are exactly the Riemannian gradient and Hessian of the respective manifolds.

orthogonal to $\boldsymbol{v}_j$. The object $\{\boldsymbol{U} \in \mathbb{R}^{n\times r} : \|\boldsymbol{u}_i\| = 1 \; \forall i\}$ is called the *oblique manifold*, which is a product space of multiple spheres. [GHJY15] showed all local minimizers of g are equivalent (i.e., signed permuted) copies of $[\boldsymbol{a}_1, \dots, \boldsymbol{a}_n]$. Moreover, g is $(1/\mathrm{poly}(n), 1/\mathrm{poly}(n), 1, 1/\mathrm{poly}(n))$-ridable.

19.2 Noisy phase synchronization and community detection

Phase synchronization concerns recovery of unit-modulus complex scalars from their relative phases. More precisely, recovering an unknown vector $\boldsymbol{z} \in \mathbb{C}_1^n$ with

$$\mathbb{C}_1^n \doteq \{\boldsymbol{z} \in \mathbb{C}^n : |z_1| = \cdots = |z_n| = 1\},$$

from noisy measurements of the form $C_{ij} = z_i\overline{z_j} + \Delta_{ij}$. This is a special version of the class of problems concerning recovery of elements from a group: given $g_i g_j^{-1} + \Delta_{ij}$ for all $i, j \in [n]$ (n possibly infinite), where all g_i's belong to a known compact group $\mathcal{G}$, recover all g_i's. Another notable example is when $\mathcal{G}$ is $\mathrm{SO}(3)$ and the problem is called *angular synchronization*. This class of problems stem from applications in signal processing, communications, computer vision, scientific imaging; see [BBS14, BCS15] for pointers.

Phase synchronization is interesting when the noise is nonzero yet controlled, which demands robust solution schemes. Turning to the optimization approach, a natural formulation (if one assumes a Gaussian noise model) is

$$\mathrm{minimize}_{\boldsymbol{x}\in\mathbb{C}_1^n} \; \|\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{C}\|_F^2,$$

where we have collected C_{ij} into a matrix $\boldsymbol{C}$. Assuming the noise is Hermitian (i.e., $\Delta_{ij} = \overline{\Delta_{ji}}$), the above formulation is equivalent to

$$\mathrm{minimize}_{\boldsymbol{x}\in\mathbb{C}_1^n} \; -\boldsymbol{x}^*\boldsymbol{C}\boldsymbol{x}. \tag{19.2.1}$$

In words, the optimization problem tries to maximize a quadratic form over products of one-dimensional circles, which is known to be NP-hard in general (Proposition 3.5 in [ZH06]). Interestingly, for the phase synchronization model, i.e., $\boldsymbol{C} = \boldsymbol{z}\boldsymbol{z}^* + \boldsymbol{\Delta}$ with Hermitian noise matrix $\boldsymbol{\Delta}$, [Bou16] recently showed that (Theorem 4) when the noise $\boldsymbol{\Delta}$ is bounded in mild sense,

> second-order necessary condition for optimality is also sufficient, and the global minimizers recover $\boldsymbol{z}$ (up to a global phase offset).

Particularly, this holds w.h.p. when the noise is i.i.d. complex Gaussians with small variance (Lemma 5). To understand the above statement, recall that second-order necessary condition asks for vanishing gradient and positive semidefinite Hessian at a point. The above statement asserts that such condition is sufficient to guarantee global optimality. In other words, at any critical points other than these verifying the condition have indefinite Hessians. Thus, [Bou16] has effectively showed that when $\boldsymbol{\Delta}$ is appropriately bounded,

> the function $-\boldsymbol{x}^* \boldsymbol{C} \boldsymbol{x}$ over $\mathbb{C}_1^n$ is a "qualitative" $\mathcal{X}$ function[2], and the global minimizers recover $\boldsymbol{z}$ (up to a global phase offset).

The real counterpart of phase synchronization is called *synchronization over* $\mathbb{Z}^2$, i.e., $\boldsymbol{z} \in \{1, -1\}^n$. In this case, an analogous formulation to (19.2.1) appears to be a hard combinatorial problem (think of MAX-CUT). Interestingly, [BBV16] showed certain *nonconvex* relaxation has a benign geometric structure. Specifically, applying the usual SDP lifting idea leads to

$$\operatorname{minimize}_{\boldsymbol{X} \in \mathbb{R}^{n \times n}} \; -\langle \boldsymbol{X}, \boldsymbol{C} \rangle \quad X_{ii} = 1, \forall\, i, \quad \boldsymbol{X} \succeq \boldsymbol{0}, \quad \operatorname{rank}(\boldsymbol{X}) = 1.$$

Dropping the rank constraint results in a convex program (SDP), which is expensive to solve for large n. The Burer-Monteiro factorization approach [BM03, BM05] suggests substituting $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{W}^\top$ for $\boldsymbol{W} \in \mathbb{R}^{n \times p}$ for $1 \le p \ll n$ such that the above relaxation is reformulated as

$$\operatorname{minimize}_{\boldsymbol{W} \in \mathbb{R}^{n \times p}} \; -\operatorname{tr}\left(\boldsymbol{W}^\top \boldsymbol{C} \boldsymbol{W}\right) \quad \left\|\boldsymbol{w}^i\right\| = 1 \;\forall\, i. \tag{19.2.2}$$

Classic results [Sha82, Bar95, Pat98] on this says problem (19.2.2) has the same optimal value as the SDP relaxation when p is large enough ($p \sim \Theta(\sqrt{n})$). Moreover, when p is set to be this scale, rank-deficient local optimizers are also global [BM05]. Surprisingly, [BBV16] showed (Theorem 4) that even $p = 2$, for the $\mathbb{Z}^2$ synchronization problem with small noise (i.e., small $\boldsymbol{\Delta}$), formulation (19.2.2) obeys

> all points verifying the second-order necessary condition are global minimizers, and any global minimizer $\boldsymbol{W}_\star$ obeys $\boldsymbol{W}_\star \boldsymbol{W}_\star^\top = \boldsymbol{z}\boldsymbol{z}^\top$.

By analogous argument to the complex case, this implies:

> the function $-\operatorname{tr}\left(\boldsymbol{W}^\top \boldsymbol{C} \boldsymbol{W}\right)$ over the oblique manifold $\left\{\boldsymbol{W} \in \mathbb{R}^{n \times 2} : \|\boldsymbol{w}_i\| = 1 \;\forall i \in [n]\right\}$ is a qualitative $\mathcal{X}$ function.

[2] Strictly speaking, our definition of $\mathcal{X}$ functions requires the function to be locally *strongly* convex around the local/global minimizers, while the Hessian being positive semidefinite is weaker than that. No matter whether their result can be strengthened in this respect, we note that we impose the strong convexity assumption instead of just convexity is for the sake of deriving concrete convergence rates for optimization algorithms. One can relax the requirement when talking of the qualitative aspect of the structure. Similar comment applies to the ensuing discussion of the real version also.

A similar result was derived in [BBV16] for the two-block community detection problem under the stochastic block model (Theorem 6).[3]

[3] Both [BBV16] and [Mon16] also contain results that characterize local optimizers in terms of their correlation with the optimizer under less stringent/general conditions on the noise.

Chapter 20

Future Directions

> There is a general principle that a stupid man can ask such questions to which one hundred wise men would not be able to answer. In accordance with this principle I shall formulate some problems.
>
> Vladimir Arnold

> To ask the right question is harder than to answer it.
>
> Georg Cantor

This thesis has been centered around the (twice continuously differentiable) $\mathcal{X}$ functions, for which all local minimizers are global, and around any saddle point there is a negative directional curvature. In other words, second-order necessary condition is also sufficient for testing global optimality. The benign $\mathcal{X}$ structure allows iterative methods that can escape from ridable saddles (and local maximizers) to convergence to a global minimizer, from arbitrary initializations. Among several choices, we have focused on the second-order trust-region method as proof of concept.

We have motivated the $\mathcal{X}$ structure with the classic Rayleigh quotient formulation for matrix eigenvector problem. To demonstrate the practical relevance and versatility of the structure, we have worked out two practical problems in great detail: complete (sparse) dictionary learning and generalized phase retrieval. In each case, we have showed a natural nonconvex formulation of the problem has the $\mathcal{X}$ structure with concrete parameters. These quantities have facilitated our development of polynomial-time convergence results for the trust-region methods adapted to each problem. The $\mathcal{X}$ structure and the resulting convergence result together have produced remarkably novel computational guarantees for both problems. Towards the end, we described two additional problems, orthogonal tensor decompositions and noisy phase synchronization/community detection, that admit nonconvex formulations with the $\mathcal{X}$ structure also.

It comes as a surprise that the $\mathcal{X}$ structure arises from dramatically diverse applications, reminiscent of convexity that dominates modeling and computation across numerous applied disciplines. The similarity goes beyond the superficial prevalence: (1) No spurious minimizers. For both convex and $\mathcal{X}$ functions, all local minimizers are also global. (2) Optimality condition. For convex functions, the first-order necessary condition is sufficient for optimality; for $\mathcal{X}$ functions, the second-order necessary condition is sufficient for optimality. (3) Optimization method. Restricted to twice continuously differentiable functions, first-order gradient descent and second-order trust-region method can serve as conceptual generic methods for optimizing convex and $\mathcal{X}$ functions, respectively (assuming proper parameter tuning and no pressure on finite-time convergence rate). So a grand question to answer is

> to what extent one can develop a theoretical and computational framework for $\mathcal{X}$ functions, parallel to convex analysis and optimization.

From practical examples we have worked out, it seems that separation of structure verification and algorithm design falls out naturally, thanks to the versatility of the (Riemannian) trust-region method (and recent attempts to provide generic convergence results thereof [BAC16]). So the central task is the analysis part, or structure verification. The success of convex analysis builds heavily on operation rules that preserve convexity, from which one can identify and construct new convex functions easily. However, it is easy to see for $\mathcal{X}$ functions, even the simple summation rule that is highly desirable fails badly. Thus, it is a standing challenge to identify practically relevant operation rules that preserve $\mathcal{X}$-ness, or to elucidate the right expectation/limitation in this regard.

An intertwined question to the above is when the $\mathcal{X}$ structure arises for particular physical problems. Generally this may be an ambitious or even wrong question to ask, because fixing one target solution, one can reverse-engineer infinitely many smooth functions with arbitrarily complex landscapes that have the target solution as its unique global minimizer. In fact, for the GPR problem we studied in Part III, under the same measurement model, a different formulation as studied in [CC15] evidently has many spurious local minimizers – possessing the $\mathcal{X}$ structure is out of question (see Figure 20.1, and compare it to Figure 12.2).

It seems valuable to narrow down the scope and ponder on the practical problems covered in this thesis that admit nonconvex formulations with the $\mathcal{X}$ structure. They all concern recovery of structured signals up to intrinsic symmetries (i.e., signs, permutations, scales, global complex phase, etc). These symmetries acting

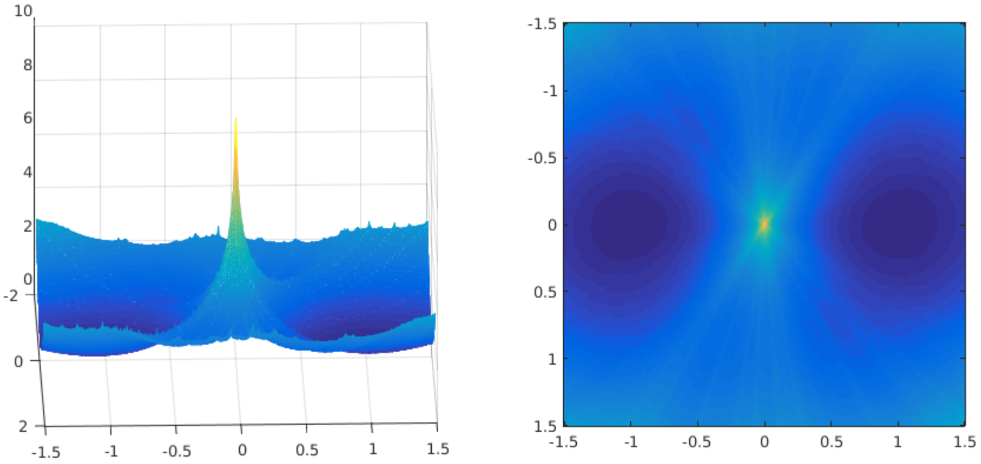


Figure 20.1: Function landscape of the GPR problem with the logarithm Poisson maximum likelihood objective. As in Figure 12.2, the measurements are i.i.d. real Gaussians and the target signal $\boldsymbol{x} = [1; 0]$ and $m \to \infty$. The objective is now $f(\boldsymbol{z}) = -\sum_{k=1}^{m} (y_k^2 \log(|\boldsymbol{a}_k^* \boldsymbol{z}|^2) - |\boldsymbol{a}_k^* \boldsymbol{z}|^2)$, which is the logarithm of the Poisson maximum likelihood. (Left) the function landscape; (Right) the same function imposed on the domain and visualized as an image. It is evident that the function landscape is very rugged and there are numerous spurious local minimizers.

on the target signals induce discrete or nonconvex target sets in the signal space. Unless the symmetries are explicitly broken, convex formulations in the original space tend to produce undesired global optimizers and hence fail in recovery. Thus, *the intrinsic symmetries favor nonconvex modeling*. Recall that $\mathcal{X}$-ness comprises two essential aspects: all local minimizers are global and all saddle points are ridable. For the former, it seems a first reasonable concern should be if the target signal is indeed a local minimizer, better still, if the function is locally convex around the target – which may not be totally unexpected when reasonable structure promoters are in use. What helps prevent the presence of spurious local minimizers may be hard to gauge, however, as we illustrated above with the GPR problem. For the latter ridability, both symmetries and randomness/genericness of the input data could play roles. Intuitively, symmetries in the space tend to cause the connecting landscape to bend down, producing negative curvatures. Moreover, recent advances in understanding random functions on manifolds (e.g., [AA+13]) prescribe that saddles of random functions (under certain randomness models) with enough smoothness tend to be ridable when the function value is remote from the optimum. Thus, *signal structures, symmetries, and randomness/genericness of input data are relevant to the $\mathcal{X}$ structure*. For recovery of structured signals, verifying local correctness (i.e., recovery as a local minimizer) is a sensible first step to take.

In view of the above discussion, striving to complete the theoretical and computational framework for $\mathcal{X}$ functions now seems premature, if not impossible. We believe a most fruitful/pragmatic way to go in the intermediate term is focusing on practical problems where nonconvex optimization is inevitable or desirable.

We discuss several concrete problems below.

For representation learning, one can easily write overcomplete dictionary learning into the form (3.1.1). To constrain the structure of the dictionary, however, one needs to work with more complicated manifolds than the sphere. Modulo the technical depth, the analytic ideas developed around the $\mathcal{X}$ structure likely generalize. Another central thread in this line, the empirical success of training deep neural networks, seems largely mysterious. Before we can completely decipher what happens underneath, understanding training neural networks with one hidden layer of nonlinearity (e.g., [JSA15]) seems a quite reasonable cutting point. Relevant problems, such as blind deconvolution (see, e.g., [LWB13]) and convolutional dictionary learning (see, e.g., [ZKTF10, BEL13]), can also be studied under the current framework.

Most tensor problems seem to ask no alternative but nonconvex approaches, as most convexity magics fail [HL13]. To date, even the most basic low-rank tensor recovery problems are still far poorly understood, as compared to the matrix counterparts [MHWG14, SRT15]. A first meaningful step would be looking at the super-structured tensors (say, low-rank "positive definite" tensors) and see how far one can carry on the gradient-descant idea, emulating similar study on positive low-rank matrices [TBSR15, ZL15, WWS15, CW15]. The completion problem arises naturally in applications, such as when modeling relations, multivariate measurements, and so on.

An important class of nonconvex problems are discrete optimization problems. Currently, the most powerful solutions to discrete problems such as clustering, community detection, submodular minimization are SDP-based convex relaxations that face significant scalability problem. It is interesting to how extensive the Burer-Monteiro factorization approach [BM03, BM05] can yield provable results, in the line of recent efforts [Bou16, BBV16]. Another thread would be considering natural analogs of gradient descent algorithms on the discrete setting and see if insights from the continuous domain can help understand practical and powerful combinatorial algorithms (e.g., on submodular minimization [CJK14]).

An alternative approach to provable nonconvex recovery starts with certain problem-dependent initializations that are hopefully already close to the target and proceeds with local convergence arguments. This strategy has been deployed on a number of problems and proves very effective and powerful (see relevant discussions in Section 3.4 and 12.4) [1]. For many practical problems of interest, however, it is often nontrivial to come up with effective initializations. Moreover, this approach has not explained why initialization-free nonconvex algorithms work well in practice. It is tempting to ask to what extent we can remove the need for

[1] The current author maintains a webpage dedicated to provable nonconvex algorithms to practical problems: http://sunju.org/research/nonconvex/. Most of the papers listed there actually take this "initialization plus local refinement" approach.

special initializations based on similar global geometric considerations we undertake here.

Moving beyond ridable saddles, practical nonconvex problems may possess saddles that are shaped by high-order derivatives, when the second-order derivatives vanish in certain directions. Identifying tractable cases and designing practical algorithms to escape from these saddles and find global optimizer is of great interest; see [AG16] for a step towards this direction. Also in view of the recent performance guarantees for first-order methods [GHJY15, LSJR16], it is natural to ask to what extent similar results can be established for higher-order structured non-ridable saddles.

Bibliography

Bibliography

[AA$^+$13] Antonio Auffinger, Gerard Ben Arous, et al. Complexity of random smooth functions on the high-dimensional sphere. *The Annals of Probability*, 41(6):4214–4247, 2013.

[AAJ$^+$13] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013.

[AAN13] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *arXiv preprint arXiv:1309.1952*, 2013.

[ABFM14] Boris Alexeev, Afonso S. Bandeira, Matthew Fickus, and Dustin G. Mixon. Phase retrieval with polarization. *SIAM Journal on Imaging Sciences*, 7(1):35–66, 2014.

[ABG07] Pierre-Antoine. Absil, Christopher G. Baker, and Kyle A. Gallivan. Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, 2007.

[ABGM13] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. *arXiv preprint arXiv:1310.6343*, 2013.

[ABGM14] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning. *arXiv preprint arXiv:1401.0579*, 2014.

[ABRS10] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[Ada16] Radosław Adamczak. A note on the sample complexity of the er-spud algorithm by spielman, wang and wright for exact recovery of sparsely used dictionaries. *arXiv preprint arXiv:1601.02049*, 2016.

[AEB06] Michal Aharon, Michael Elad, and Alfred M Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear algebra and its applications*, 416(1):48–67, 2006.

[AG16] Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *arXiv preprint arXiv:1602.05908*, 2016.

[AGH$^+$14] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[AGJ14a] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models. *arXiv preprint arXiv:1411.1488*, 2014.

[AGJ14b] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.

[AGKM12] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM, 2012.

[AGM13] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. *arXiv preprint arXiv:1308.6273*, 2013.

[AGMM15] Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*, 2015.

[AGMS12] Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ICA with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In *Advances in Neural Information Processing Systems*, pages 2375–2383, 2012.

[AHJK13] A Anandkumar, D Hsu, M Janzamin, and SM Kakade. When are overcomplete topic models identifiable. *Uniqueness of Tensor Tucker Decompositions with Structured Sparsity. ArXiv*, 1308, 2013.

[AJSN15] Animashree Anandkumar, Prateek Jain, Yang Shi, and Uma Naresh Niranjan. Tensor vs matrix methods: Robust tensor decomposition under block sparse perturbations. *arXiv preprint arXiv:1510.04747*, 2015.

[AL12] Miguel F Anjos and Jean B Lasserre. *Introduction to semidefinite, conic and polynomial optimization*. Springer, 2012.

[ALMT14] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, page iau005, 2014.

[AM12] P.-A. Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

[AMS09] Pierre-Antoine. Absil, Robert Mahoney, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[ARR14] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *Information Theory, IEEE Transactions on*, 60(3):1711–1732, 2014.

[Bac10] Francis R Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, pages 118–126, 2010.

[BAC16] Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Complexity of nonconvex optimization on manifolds. In preparation, 2016.

[Bal10] Radu V. Balan. On signal reconstruction from its spectrogram. In *Information Sciences and Systems (CISS), 44th Annual Conference on*, pages 1–4. IEEE, 2010.

[Bar95] Alexander I. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(2):189–202, 1995.

[BBCE09] Radu Balan, Bernhard G. Bodmann, Peter G. Casazza, and Dan Edidin. Painless reconstruction from magnitudes of frame coefficients. *Journal of Fourier Analysis and Applications*, 15(4):488–501, 2009.

[BBS14] Afonso S Bandeira, Nicolas Boumal, and Amit Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *arXiv preprint arXiv:1411.3272*, 2014.

[BBV16] Afonso S Bandeira, Nicolas Boumal, and Vladislav Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. *arXiv preprint arXiv:1602.04426*, 2016.

[BCE06] Radu Balana, Pete Casazzab, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345 – 356, 2006.

[BCJ13] Chenglong Bao, Jian-Feng Cai, and Hui Ji. Fast sparsity-based orthogonal dictionary learning for image restoration. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3384–3391. IEEE, 2013.

[BCS15] Afonso S Bandeira, Yutong Chen, and Amit Singer. Non-unique games over compact groups and orientation estimation in cryo-em. *arXiv preprint arXiv:1505.03840*, 2015.

[BDP$^+$07] Oliver Bunk, Ana Diaz, Franz Pfeiffer, Christian David, Bernd Schmitt, Dillip K. Satapathy, and J. Friso van der Veen. Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. *Acta Crystallographica Section A*, 63(4):306–314, Jul. 2007.

[BEGFB94] Stephen P Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear matrix inequalities in system and control theory*, volume 15. SIAM, 1994.

[BEL13] Hilton Bristow, Anders Eriksson, and Simon Lucey. Fast convolutional sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 391–398, 2013.

[Ber99] Dimitri P. Bertsekas. *Nonlinear programming*. Athena scientific, 1999.

[BJQS14] Chenglong Bao, Hui Ji, Yuhui Quan, and Zuowei Shen. L0 norm based dictionary learning by proximal methods with global convergence. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3858–3865. IEEE, 2014.

[BJS14] Chenglong Bao, Hui Ji, and Zuowei Shen. Convergence analysis for iterative data-driven tight frame construction scheme. *Applied and Computational Harmonic Analysis*, 2014.

[BKS13a] Afonso S Bandeira, Christopher Kennedy, and Amit Singer. Approximating the little grothendieck problem over the orthogonal and unitary groups. *arXiv preprint arXiv:1308.5207*, 2013.

[BKS13b] Boaz Barak, Jonathan Kelner, and David Steurer. Rounding sum-of-squares relaxations. *arXiv preprint arXiv:1312.6652*, 2013.

[BKS14] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. *arXiv preprint arXiv:1407.1543*, 2014.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[BM03] Samuel Burer and Renato D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[BM05] Samuel Burer and Renato D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[BMAS14] Nicolas Boumal, Bamdev Mishra, P.-A. Absil, and Rodolphe Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014.

[BN16] Jarosław Błasiok and Jelani Nelson. An improved analysis of the er-spud dictionary learning algorithm. *arXiv preprint arXiv:1602.05719*, 2016.

[Bou16] Nicolas Boumal. Nonconvex phase synchronization. *arXiv preprint arXiv:1601.06114*, 2016.

[BQJ14] Chenglong Bao, Yuhui Quan, and Hui Ji. A convergent incoherent dictionary learning algorithm for sparse coding. In *Computer Vision–ECCV 2014*, pages 302–316. Springer, 2014.

[BR13] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, 2013.

[BR14] Jop Briët and Oded Regev. Tight hardness of the non-commutative grothendieck problem. *arXiv preprint arXiv:1412.4413*, 2014.

[BS14] Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *arXiv preprint arXiv:1404.5236*, 2014.

[BST14] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[BT89] Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

[Bur12] Samuel Burer. Copositive programming. In *Handbook on semidefinite, conic and polynomial optimization*, pages 201–218. Springer, 2012.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[BWY14] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *arXiv preprint arXiv:1408.2156*, 2014.

[BZ87] Andrew Blake and Andrew Zisserman. *Visual reconstruction*, volume 2. MIT press Cambridge, 1987.

[Can02] Emmanuel J. Candès. New ties between computational harmonic analysis and approximation theory. *Approximation Theory X*, pages 87–153, 2002.

[Can14] Emmanuel J. Candès. Mathematics of sparsity (and few other things). In *Proceedings of the International Congress of Mathematicians, Seoul, South Korea*, 2014.

[CC15] Yuxin Chen and Emmanuel J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *arXiv preprint arXiv:1505.05114*, 2015.

[CESV13] Emmanuel J. Candès, Yonina C. Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1), 2013.

[CGT00] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust-region Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[CJK14] Deeparnab Chakrabarty, Prateek Jain, and Pravesh Kothari. Provable submodular minimization using wolfe's algorithm. In *Advances in Neural Information Processing Systems*, pages 802–809, 2014.

[CL14] Emmanuel J. Candès and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2014.

[CLM15] T. Tony Cai, Xiaodong Li, and Zongming Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *arXiv preprint arXiv:1506.03382*, 2015.

[CLMW11] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[CLS15a] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2015.

[CLS15b] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007, April 2015.

[CM14] Sunav Choudhary and Urbashi Mitra. Identifiability scaling laws in bilinear inverse problems. *arXiv preprint arXiv:1402.2637*, 2014.

[CMP11] Anwei Chai, Miguel Moscoso, and George Papanicolaou. Array imaging using intensity-only measurements. *Inverse Problems*, 27(1):015005, 2011.

[Com94] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[Cor06] John V. Corbett. The pauli problem, state reconstruction and quantum-real numbers. *Reports on Mathematical Physics*, 57(1):53–68, 2006.

[CRPW12] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[CSV13] 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.

[CW15] Yudong Chen and Martin J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[DeV98] Ronald A. DeVore. Nonlinear approximation. *Acta numerica*, 7:51–150, 1998.

[DeV09] Ronald A DeVore. Nonlinear approximation and its applications. In *Multiscale, Nonlinear and Adaptive Approximation*, pages 169–201. Springer, 2009.

[DF87] Chris Dainty and James R. Fienup. Phase retrieval and image reconstruction for astronomy. *Image Recovery: Theory and Application*, pages 231–275, 1987.

[DGM13] David L Donoho, Matan Gavish, and Andrea Montanari. The phase transition of matrix recovery from gaussian measurements matches the minimax mse of matrix denoising. *Proceedings of the National Academy of Sciences*, 110(21):8405–8410, 2013.

[DH14] Laurent Demanet and Paul Hand. Scaling law for recovering the sparsest element in a subspace. *Information and Inference*, 3(4):295–309, 2014.

[DJ94] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[Don95] David L Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995.

[DT09] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[Due10] Lutz Duembgen. Bounding standard gaussian tail probabilities. *arXiv preprint arXiv:1012.2063*, 2010.

[DVDD98] David L. Donoho, Martin Vetterli, Ronald A. DeVore, and Ingrid Daubechies. Data compression and harmonic analysis. *Information Theory, IEEE Transactions on*, 44(6):2435–2476, 1998.

[EAS98] 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.

[Ela10] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

[EW15] Armin Eftekhari and Michael B. Wakin. Greed is super: A fast algorithm for super-resolution. *arXiv preprint arXiv:1511.03385*, 2015.

[Fie82] James R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, Aug 1982.

[FJK96] Alan Frieze, Mark Jerrum, and Ravi Kannan. Learning linear transformations. In *focs*, page 359. IEEE, 1996.

[Fol99] Gerald B Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2nd edition, 1999.

[FR13] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

[FW04] Charles Fortin and Henry Wolkowicz. The trust region subproblem and semidefinite programming. *Optimization methods and software*, 19(1):41–67, 2004.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[GJB$^+$13] Remi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, and Matthias Seibert. Sample complexity of dictionary learning and other matrix factorizations. *arXiv preprint arXiv:1312.3790*, 2013.

[GJB14] Rémi Gribonval, Rodolphe Jenatton, and Francis Bach. Sparse and spurious: dictionary learning with noise and outliers. *arXiv preprint arXiv:1407.5155*, 2014.

[GKK13] David Gross, Felix Krahmer, and Richard Kueng. A partial derandomization of phaselift using spherical designs. *arXiv preprint arXiv:1310.2267*, 2013.

[GN10] Lee-Ad Gottlieb and Tyler Neylon. Matrix sparsification and the sparse null space problem. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 205–218. Springer, 2010.

[Gol80] Donald Goldfarb. Curvilinear path steplength algorithms for minimization which use directions of negative curvature. *Mathematical programming*, 18(1):31–40, 1980.

[GS72] R. W. Gerchberg and W. Owen Saxton. A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[GS10] Rémi Gribonval and Karin Schnass. Dictionary identification - sparse matrix-factorization via ℓ^1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

[GVL12] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[GW95] Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[GW11] Quan Geng and John Wright. On the local correctness of ℓ^1-minimization for dictionary learning. Submitted to *IEEE Transactions on Information Theory*, 2011. Preprint: http://www.columbia.edu/~jw2966.

[Har60] Theodore E. Harris. A lower bound for the critical probability in a certain percolation process. *Mathematical Proceedings of the Cambridge Philosophical Society*, 56(01):13–20, 1960.

[Har14] Moritz Hardt. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 651–660. IEEE, 2014.

[Hig08] Nicholas J. Higham. *Functions of Matrices*. Society for Industrial and Applied Mathematics, 2008.

[HK14] Elad Hazan and Tomer Koren. A linear-time algorithm for trust region problems. *arXiv preprint arXiv:1401.6757*, 2014.

[HKO01] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons., 2001.

[HL13] Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

[HMG94] Uwe Helmke, John B Moore, and Würzburg Germany. *Optimization and dynamical systems*. Springer-Verlag London, 1994.

[HMW13] Teiko Heinosaari, Luca Mazzarella, and Michael M. Wolf. Quantum tomography under prior information. *Communications in Mathematical Physics*, 318(2):355–374, 2013.

[HO00] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[HP13] Reiner Horst and Panos M Pardalos. *Handbook of global optimization*, volume 2. Springer Science & Business Media, 2013.

[HPVT00] Reiner Horst, Panos M Pardalos, and Nguyen Van Thoai. *Introduction to global optimization*. Springer Science & Business Media, 2000.

[HS11] Christopher Hillar and Friedrich T Sommer. When can dictionary learning uniquely recover sparse data from subsamples? *arXiv preprint arXiv:1106.3616*, 2011.

[HSSS15] Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Speeding up sum-of-squares for tensor decomposition and planted sparse vectors. *arXiv preprint arXiv:1512.02337*, 2015.

[HUL93a] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer-Verlag, New York, 1993.

[HUL93b] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms II: Advanced theory and bundle methods*, volume 306. Springer-Verlag, New York, 1993.

[HW14] Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. In *Proceedings of The 27th Conference on Learning Theory*, pages 638–678, 2014.

[Hyv99] Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10(3):626–634, 1999.

[JEH15] Kishore Jaganathan, Yonina C. Eldar, and Babak Hassibi. Phase retrieval: An overview of recent developments. *arXiv preprint arXiv:1510.07713*, 2015.

[JJKN15] Prateek Jain, Chi Jin, Sham M. Kakade, and Praneeth Netrapalli. Computing matrix squareroot via non convex local search. *arXiv preprint arXiv:1507.05854*, 2015.

[JN14] Prateek Jain and Praneeth Netrapalli. Fast exact matrix completion with finite samples. *arXiv preprint arXiv:1411.1087*, 2014.

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, pages 665–674. ACM, 2013.

[JO14] Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.

[JOH13] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Sparse phase retrieval: Convex algorithms and limitations. In *Proceedings of IEEE International Symposium on Information Theory*, pages 1022–1026. IEEE, 2013.

[JSA15] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *CoRR abs/1506.08473*, 2015.

[KB09] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[KD09] Ken Kreutz-Delgado. The complex gradient operator and the $\mathbb{CR}$-calculus. *arXiv preprint arXiv:0906.4835*, 2009.

[KMO10] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.

[Las07] Jean B Lasserre. A sum of squares approximation of nonnegative polynomials. *SIAM review*, 49(4):651–669, 2007.

[LJ15] Kiryung Lee and Marius Junge. RIP-like properties in subsampled blind deconvolution. *arXiv preprint arXiv:1511.06146*, 2015.

[LLJB15] Kiryung Lee, Yanjun Li, Marius Junge, and Yoram Bresler. Blind recovery of sparse signals from subsampled convolution. *arXiv preprint arXiv:1511.06149*, 2015.

[Loh15] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *arXiv preprint arXiv:1501.00312*, 2015.

[LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.

[LSSS14] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

[LV13] Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

[LV15] Kyle Luh and Van Vu. Dictionary learning with few samples and matrix concentration. *arXiv preprint arXiv:1503.08854*, 2015.

[LW11] Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

[LW13] Po-Ling Loh and Martin J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

[LW14] Po-Ling Loh and Martin J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *arXiv preprint arXiv:1412.5632*, 2014.

[LWB13] Kiryung Lee, Yihong Wu, and Yoram Bresler. Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization. *arXiv preprint arXiv:1312.0525*, 2013.

[Mat02] Yukio Matsumoto. *An introduction to Morse theory*, volume 208. American Mathematical Soc., 2002.

[MBP14] Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

[MFI15a] Hossein Mobahi and John W Fisher III. On the link between gaussian homotopy continuation and convex envelopes. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 43–56. Springer, 2015.

[MFI15b] Hossein Mobahi and John W Fisher III. A theoretical analysis of optimization by gaussian continuation. In *AAAI*, pages 1205–1211, 2015.

[MG13] Nishant Mehta and Alexander G. Gray. Sparsity-based generalization bounds for predictive sparse coding. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 28(1):36–44, 2013.

[MHWG14] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved convex relaxations for tensor recovery. *Journal of Machine Learning Research*, 1:1–48, 2014.

[MIJ$^+$02] Jianwei Miao, Tetsuya Ishikawa, Bart Johnson, Erik H. Anderson, Barry Lai, and Keith O. Hodgson. High resolution 3D X-Ray diffraction microscopy. *Phys. Rev. Lett.*, 89(8):088303, Aug 2002.

[Mil90] R. P. Millane. Phase retrieval in crystallography and optics. *Journal of the Optical Society of America A*, 7(3):394–411, Mar 1990.

[MK87] Katta G. Murty and Santosh N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

[Mon16] Andrea Montanari. A grothendieck-type inequality for local maxima. *arXiv preprint arXiv:1603.04064*, 2016.

[MP10a] Jianwei Ma and Gerlind Plonka. A review of curvelets and recent applications. *IEEE Signal Processing Magazine*, 27(2):118–133, 2010.

[MP10b] Andreas Maurer and Massimiliano Pontil. K-dimensional coding schemes in hilbert spaces. *Information Theory, IEEE Transactions on*, 56(11):5839–5846, 2010.

[MS83] Jorge J. Moré and Danny C. Sorensen. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4(3):553–572, 1983.

[MT14] Michael B McCoy and Joel A Tropp. Sharp recovery bounds for convex demixing, with applications. *Foundations of Computational Mathematics*, 14(3):503–567, 2014.

[Nic11] Liviu Nicolaescu. *An invitation to Morse theory*. Springer Science & Business Media, second edition, 2011.

[NJS13] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

[NNS$^+$14] Praneeth Netrapalli, Uma Naresh. Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[NP06] Yurii Nesterov and Boris T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[NP13] Behnam Neyshabur and Rina Panigrahy. Sparse matrix factorization. *arXiv preprint arXiv:1311.3315*, 2013.

[NW06] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer, 2006.

[NWY00] Yuri Nesterov, Henry Wolkowicz, and Yinyu Ye. Semidefinite programming relaxations of nonconvex quadratic optimization. In *Handbook of semidefinite programming*, pages 361–419. Springer, 2000.

[NYWR09] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[OCPB13] Brendan OâĂŹDonoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *arXiv preprint arXiv*, 1312, 2013.

[OF96] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[OF97] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[OH10] Samet Oymak and Babak Hassibi. New null space results and recovery thresholds for matrix rank minimization. *arXiv preprint arXiv:1011.6326*, 2010.

[OJF$^+$12] Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C. Eldar, and Babak Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753*, 2012.

[OYDS12] Henrik Ohlsson, Allen Y. Yang, Roy Dong, and S. Shankar Sastry. CPRL – An extension of compressive sensing to the phase retrieval problem. In *Advances in Neural Information Processing Systems*, 2012.

[OYDS13] Henrik Ohlsson, Allen Y. Yang, Roy Dong, and S. Shankar Sastry. Compressive phase retrieval from squared output measurements via semidefinite programming. *arXiv preprint arXiv:1111.6323*, 2013.

[OYVS13] Henrik Ohlsson, Allen Y. Yang, Michel Verhaegen, and S. Shankar Sastry. Quadratic basis pursuit. *arXiv preprint arXiv:1301.7002*, 2013.

[Par03] Pablo A Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical programming*, 96(2):293–320, 2003.

[Pat98] Gábor Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research*, 23(2):339–358, 1998.

[QSW14] Qing Qu, Ju Sun, and John Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. In *Advances in Neural Information Processing Systems*, pages 3401–3409, 2014.

[Rap97] Tamás Rapcsák. *Smooth nonlinear optimization in $\mathbb{R}^n$*, volume 19. Springer Science & Business Media, 1997.

[Rei65] H. Reichenbach. *Philosophic foundations of quantum mechanics*. University of California Press, 1965.

[Rob93] W. Harrison Robert. Phase problem in crystallography. *Journal of the Optical Society of America A*, 10(5):1046–1055, 1993.

[RW97] Franz Rendl and Henry Wolkowicz. A semidefinite framework for trust region subproblems with applications to large scale minimization. *Mathematical Programming*, 77(1):273–299, 1997.

[SA14a] Hanie Sedghi and Anima Anandkumar. Provable methods for training neural networks with sparse connectivity. *arXiv preprint arXiv:1412.2693*, 2014.

[SA14b] Hanie Sedghi and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. *arXiv preprint arXiv:1412.3046*, 2014.

[SA14c] Hanie Sedghi and Animashree Anandkumar. Provable tensor methods for learning mixtures of classifiers. *arXiv preprint arXiv:1412.3046*, 2014.

[SBE14] Yoav Shechtman, Amir Beck, and Yonina C. Eldar. GESPAR: Efficient phase retrieval of sparse signals. *Signal Processing, IEEE Transactions on*, 62(4):928–938, Feb 2014.

[Sch14a] Karin Schnass. Local identification of overcomplete dictionaries. *arXiv preprint arXiv:1401.6354*, 2014.

[Sch14b] Karin Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying k-svd. *Applied and Computational Harmonic Analysis*, 37(3):464–491, 2014.

[Sch15] Karin Schnass. Convergence radius and sample complexity of itkm algorithms for dictionary learning. *arXiv preprint arXiv:1503.07027*, 2015.

[SEC$^+$15] Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry N. Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *Signal Processing Magazine, IEEE*, 32(3):87–109, May 2015.

[Sha82] Alexander Shapiro. Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika*, 47(2):187–199, 1982.

[SL14] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *arXiv preprint arXiv:1411.8003*, 2014.

[SLLC15] Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *arXiv preprint arXiv:1502.01425*, 2015.

[Sol14] Mahdi Soltanolkotabi. *Algorithms and theory for clustering and nonconvex quadratic programming*. PhD thesis, Stanford University, 2014.

[SQW15a] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere. *arXiv preprint arXiv:1504.06785*, 2015.

[SQW15b] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[SQW16] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retreival. *arXiv preprint arXiv:1602.06664*, 2016.

[SRO15] Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *The 32nd International Conference on Machine Learning*, volume 37, pages 2332–2341, 2015.

[SRT15] Parikshit Shah, Nikhil Rao, and Gongguo Tang. Optimal low-rank tensor recovery from separable measurements: Four contractions suffice. *arXiv preprint arXiv:1505.04085*, 2015.

[SS90] Gilbert W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic press, 1990.

[SWW12] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

[TBSR15] Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

[TC96] Neil A. Thacker and Timothy F. Cootes. Vision through optimization. *BMVC Tutorial Notes*, 1996.

[Tem03] Vladimir N Temlyakov. Nonlinear methods of approximation. *Foundations of Computational Mathematics*, 3(1):33–107, 2003.

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[Tro15a] Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

[Tro15b] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[Tse01] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[Udr94] Constantin Udriste. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, 1994.

[Vav09] Stephen A. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.

[Ver12] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012.

[VMB11] Daniel Vainsencher, Shie Mannor, and Alfred M. Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12(23):3259–3281, November 2011.

[VX14] Vladislav Voroninski and Zhiqiang Xu. A strong restricted isometry property, with an application to phaseless compressed sensing. *arXiv preprint arXiv:1404.3811*, 2014.

[Wal63] Adriaan Walther. The question of phase retrieval in optics. *Journal of Modern Optics*, 10(1):41–49, 1963.

[WCCL15] Ke Wei, Jian-Feng Cai, Tony F. Chan, and Shingyu Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *arXiv preprint arXiv:1511.01562*, 2015.

[WdM15] Irène Waldspurger, Alexandre d`Aspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[WGNL14] Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.

[WLL14] Zhaoran Wang, Huanran Lu, and Han Liu. Nonconvex statistical optimization: minimax-optimal sparse pca in polynomial time. *arXiv preprint arXiv:1408.5352*, 2014.

[WWS15] Chris D. White, Rachel Ward, and Sujay Sanghavi. The local convexity of solving quadratic equations. *arXiv preprint arXiv:1506.07868*, 2015.

[WY15] Siqi Wu and Bin Yu. Local identifiability of ℓ_1-minimization dictionary learning: a sufficient and almost necessary condition. *arXiv preprint arXiv:1505.04363*, 2015.

[YCS13] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. *arXiv preprint arXiv:1310.3745*, 2013.

[YZ03] Yinyu Ye and Shuzhong Zhang. New results on quadratic minimization. *SIAM Journal on Optimization*, 14(1):245–267, 2003.

[ZH06] Shuzhong Zhang and Yongwei Huang. Complex quadratic optimization and semidefinite programming. *SIAM Journal on Optimization*, 16(3):871–890, 2006.

[ZKTF10] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2528–2535. IEEE, 2010.

[ZL15] Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *arXiv preprint arXiv:1506.06081*, 2015.

[ZP01] Michael Zibulevsky and Barak Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882, 2001.

Appendices

Appendix A

Concentration Inequalities

Various quantities involved in this thesis are probabilistic in nature; establishing their properties relies heavily on the fact that under technical conditions they do not deviate much from their expectations. The latter fact is formalized as concentration inequalities. In this chapter, we record concentration inequalities we use frequently in this thesis; systematic treatment of concentration inequalities can be found in, e.g., [BLM13], and [Tro15b] which deals with especially matrix-valued random variables.

Lemma A.1 (Moment-Control Bernstein's Inequality for Scalar RVs, Theorem 2.10 of [FR13]) *Let X_1, ..., X_p be i.i.d. real-valued random variables. Suppose that there exist positive numbers R and σ^2 such that*

$$\mathbb{E}\left[|X_k|^m\right] \le \frac{m!}{2}\sigma^2 R^{m-2}, \quad \textit{for all integers } m \ge 2.$$

Let $S \doteq \frac{1}{p}\sum_{k=1}^{p} X_k$, then for all $t > 0$, it holds that

$$\mathbb{P}\left[|S - \mathbb{E}\left[S\right]| \ge t\right] \le 2\exp\left(-\frac{pt^2}{2\sigma^2 + 2Rt}\right). \tag{A.0.1}$$

Lemma A.2 (Moment-Control Bernstein's Inequality for Matrix RVs, Theorem 6.2 of [Tro12]) *Let $\boldsymbol{X}_1$, ..., $\boldsymbol{X}_p$ be i.i.d. $d \times d$ random, symmetric matrices. Suppose there exist positive numbers R and σ^2 such that*

$$\mathbb{E}\left[\boldsymbol{X}_k^m\right] \preceq \frac{m!}{2}\sigma^2 R^{m-2}\boldsymbol{I} \textit{ and } -\mathbb{E}\left[\boldsymbol{X}_k^m\right] \preceq \frac{m!}{2}\sigma^2 R^{m-2}\boldsymbol{I}\ , \textit{ for all integers } m \ge 2.$$

Let $\boldsymbol{S} \doteq \frac{1}{p}\sum_{k=1}^{p} \boldsymbol{X}_k$, then for all $t > 0$, it holds that

$$\mathbb{P}\left[\|\boldsymbol{S} - \mathbb{E}\left[\boldsymbol{S}\right]\| \ge t\right] \le 2d\exp\left(-\frac{pt^2}{2\sigma^2 + 2Rt}\right). \tag{A.0.2}$$

Proving this lemma requires some modification to the original proof of Theorem 6.2 in [Tro12]. We record it here for the sake of completeness.

Proof Let us define $\boldsymbol{S}_p = \sum_{k=1}^{p} \boldsymbol{X}_k$, by Proposition 3.1 of [Tro12], we have

$$\mathbb{P}\left[\lambda_{\max}\left(\boldsymbol{S}_p - \mathbb{E}\left[\boldsymbol{S}_p\right]\right) \geq t\right] \;\leq\; \inf_{t>0} e^{-\theta t}\mathbb{E}\left[\operatorname{tr}\exp\left(\theta \boldsymbol{S}_p - \theta\mathbb{E}\left[\boldsymbol{S}_p\right]\right)\right], \tag{A.0.3}$$

To proceed, notice that

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{tr}\exp\left(\theta \boldsymbol{S}_p - \theta\mathbb{E}\left[\boldsymbol{S}_p\right]\right)\right]\\
=\,&\mathbb{E}_{\boldsymbol{S}_{p-1}}\mathbb{E}_{\boldsymbol{X}_p}\left[\operatorname{tr}\exp\left(\theta\left(\boldsymbol{S}_{p-1} - \mathbb{E}\left[\boldsymbol{S}_{p-1}\right]\right) + \theta\boldsymbol{X}_p - \theta\mathbb{E}\left[\boldsymbol{X}_p\right]\right)\right]\\
\leq\,&\mathbb{E}_{\boldsymbol{S}_{p-1}}\left[\operatorname{tr}\exp\left(\theta(\boldsymbol{S}_{p-1} - \mathbb{E}\left[\boldsymbol{S}_{p-1}\right]) + \log\left(\mathbb{E}\left[e^{\theta\boldsymbol{X}_p}\right]\right) - \theta\mathbb{E}\left[\boldsymbol{X}_p\right]\right)\right]\\
\leq\,&\mathbb{E}_{\boldsymbol{S}_{p-1}}\left[\operatorname{tr}\exp\left(\theta(\boldsymbol{S}_{p-1} - \mathbb{E}\left[\boldsymbol{S}_{p-1}\right]) + \mathbb{E}\left[e^{\theta\boldsymbol{X}_p}\right] - \boldsymbol{I} - \theta\mathbb{E}\left[\boldsymbol{X}_p\right]\right)\right]\\
=\,&\mathbb{E}_{\boldsymbol{S}_{p-1}}\left[\operatorname{tr}\exp\left(\theta(\boldsymbol{S}_{p-1} - \mathbb{E}\left[\boldsymbol{S}_{p-1}\right]) + \sum_{\ell=2}^{\infty}\frac{\theta^\ell\mathbb{E}\left[\boldsymbol{X}_k^\ell\right]}{\ell!}\right)\right]
\end{aligned}$$

where at the third line we have used the result of Corollary 3.3 of [Tro12], i.e., $\mathbb{E}\left[\operatorname{tr}\exp\left(\boldsymbol{H} + \boldsymbol{X}\right)\right] \leq \operatorname{tr}\exp\left(\boldsymbol{H} + \log\left(\mathbb{E}\left[e^{\boldsymbol{X}}\right]\right)\right)$ for any fixed $\boldsymbol{H}$ and random, symmetric $\boldsymbol{X}$, at the fourth we have used the fact that $\log\boldsymbol{X} \preceq \boldsymbol{X} - \boldsymbol{I}$ for any $\boldsymbol{X} \succ \boldsymbol{0}$ (as $\log u \leq u - 1$ for any $u > 0$ and transfer rule applies here), and the last line relies on exchange of infinite summation and expectation, justified as $\boldsymbol{X}_p$ has a bounded spectral radius. By repeating the argument backwards for $\boldsymbol{X}_{p-1}, \cdots, \boldsymbol{X}_1$, we get

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{tr}\exp\left(\theta \boldsymbol{S}_p - \theta\mathbb{E}\left[\boldsymbol{S}_p\right]\right)\right]\\
\leq\;&\operatorname{tr}\exp\left(p\sum_{\ell=2}^{\infty}\frac{\theta^\ell\mathbb{E}\left[\boldsymbol{X}_k^\ell\right]}{\ell!}\right) \leq \operatorname{tr}\exp\left(p\sum_{\ell=2}^{p}\frac{\theta^\ell\sigma^2R^{\ell-2}}{2}\boldsymbol{I}\right)\\
\leq\;& d\left\|\exp\left(p\sum_{\ell=2}^{p}\frac{\theta^\ell\sigma^2R^{\ell-2}}{2}\boldsymbol{I}\right)\right\| \leq d\exp\left(\frac{p\theta^2\sigma^2}{2(1-\theta R)}\right),
\end{aligned} \tag{A.0.4}$$

where we used the fact that $\mathbb{E}\left[\boldsymbol{X}_i^m\right] \preceq \frac{m!}{2}\sigma^2R^{m-2}\boldsymbol{I}$ in (A.0.2) and restrict $\theta < \frac{1}{R}$. Combining the results in (A.0.3) and (A.0.4), we have

$$\mathbb{P}\left[\lambda_{\max}\left(\boldsymbol{S}_p - \mathbb{E}\left[\boldsymbol{S}_p\right]\right) \geq t\right] \;\leq\; d\inf_{\theta<1/R}\exp\left(\frac{p\theta^2\sigma^2}{2(1-\theta R)} - \theta t\right) \tag{A.0.5}$$

by taking $\theta = t/(p\sigma^2 + Rt) < 1/R$, we obtain

$$\mathbb{P}\left[\lambda_{\max}\left(\boldsymbol{S}_p - \mathbb{E}\left[\boldsymbol{S}_p\right]\right) \geq t\right] \leq d\exp\left(-\frac{t^2}{2p\sigma^2 + 2Rt}\right). \tag{A.0.6}$$

Considering $\boldsymbol{X}_k' = -\boldsymbol{X}_k$ and repeating the above argument, we can similarly obtain

$$\mathbb{P}\left[\lambda_{\min}\left(\boldsymbol{S}_p - \mathbb{E}\left[\boldsymbol{S}_p\right]\right) \le -t\right] \le d \exp\left(-\frac{t^2}{2p\sigma^2 + 2Rt}\right). \tag{A.0.7}$$

Putting the above bounds together, we have

$$\mathbb{P}\left[\left\|\boldsymbol{S}_p - \mathbb{E}\left[\boldsymbol{S}_p\right]\right\| \ge t\right] \;\le\; 2d \exp\left(-\frac{t^2}{2p\sigma^2 + 2Rt}\right). \tag{A.0.8}$$

We obtain the claimed bound by substituting $\boldsymbol{S}_p = p\boldsymbol{S}$ and simplifying the resulting expressions. ■

Corollary A.3 (Moment-Control Bernstein's Inequality for Vector RVs) *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p \in \mathbb{R}^d$ be i.i.d. random vectors. Suppose there exist some positive number R and σ^2 such that*

$$\mathbb{E}\left[\left\|\boldsymbol{x}_k\right\|^m\right] \;\le\; \frac{m!}{2}\sigma^2 R^{m-2}, \quad \textit{for all integers } m \ge 2.$$

Let $\boldsymbol{s} = \frac{1}{p}\sum_{k=1}^{p} \boldsymbol{x}_k$, then for any $t > 0$, it holds that

$$\mathbb{P}\left[\left\|\boldsymbol{s} - \mathbb{E}\left[\boldsymbol{s}\right]\right\| \ge t\right] \;\le\; 2(d+1) \exp\left(-\frac{pt^2}{2\sigma^2 + 2Rt}\right). \tag{A.0.9}$$

Proof To obtain the result, we apply the matrix Bernstein inequality in Lemma A.2 to a suitable embedding of the random vectors $\{\boldsymbol{x}_k\}_{k=1}^p$. For any $k \in [p]$, define the symmetric matrix

$$\boldsymbol{X}_k = \begin{bmatrix} 0 & \boldsymbol{x}_k^* \\ \boldsymbol{x}_k & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}.$$

Then it holds that

$$\boldsymbol{X}_k^{2\ell+1} = \left\|\boldsymbol{x}_k\right\|_2^{2\ell} \begin{bmatrix} 0 & \boldsymbol{x}_k^* \\ \boldsymbol{x}_k & \boldsymbol{0} \end{bmatrix}, \; \boldsymbol{X}_k^{2\ell+2} = \left\|\boldsymbol{x}_k\right\|^{2\ell} \begin{bmatrix} \left\|\boldsymbol{x}_k\right\|^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{x}_k\boldsymbol{x}_k^* \end{bmatrix}, \; \text{for all integers } \ell \ge 0.$$

Using the fact that

$$\boldsymbol{x}_k\boldsymbol{x}_k^* \preceq \left\|\boldsymbol{x}_k\right\|^2 \boldsymbol{I}, \quad \left\|\boldsymbol{X}_k\right\| = \sqrt{\left\|\boldsymbol{X}_k^2\right\|} = \left\|\boldsymbol{x}_k\right\| \Longrightarrow -\left\|\boldsymbol{x}_k\right\| \boldsymbol{I} \preceq \boldsymbol{X}_k \preceq \left\|\boldsymbol{x}_k\right\| \boldsymbol{I},$$

and combining the above expressions for $\boldsymbol{X}_k^{2\ell+1}$ and $\boldsymbol{X}_k^{2\ell+2}$, we obtain

$$\mathbb{E}\left[\boldsymbol{X}_k^m\right], -\mathbb{E}\left[\boldsymbol{X}_k^m\right] \preceq \mathbb{E}\left[\left\|\boldsymbol{x}_k\right\|_2^m\right] \boldsymbol{I} \preceq \frac{m!}{2}\sigma^2 R^{m-2}\boldsymbol{I}, \quad \text{for all integers } m \ge 2, \tag{A.0.10}$$

Let $\boldsymbol{S} = \frac{1}{p}\sum_{k=1}^{p} \boldsymbol{X}_k$, noting that

$$\left\| \boldsymbol{S} - \mathbb{E}\left[\boldsymbol{S}\right] \right\| = \left\| \boldsymbol{s} - \mathbb{E}\left[\boldsymbol{s}\right] \right\|, \tag{A.0.11}$$

and applying Lemma A.2, we complete the proof. ■

Lemma A.4 (Hoeffding-type Inequality, Proposition 5.10 of [Ver12]) *Let $X_1, \cdots, X_N$ be independent centered sub-Gaussian random variables, and let $K = \max_i \left\|X_i\right\|_{\psi_2}$, where the sub-Gaussian norm*

$$\left\|X_i\right\|_{\psi_2} \doteq \sup_{p\geq 1} p^{-1/2} \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}. \tag{A.0.12}$$

Then for every $\boldsymbol{b} = [b_1; \cdots; b_N] \in \mathbb{C}^N$ and every $t \geq 0$, we have

$$\mathbb{P}\left(\left| \sum_{k=1}^{N} b_k X_k \right| \geq t \right) \leq e \exp\left(-\frac{ct^2}{K^2 \left\|\boldsymbol{b}\right\|_2^2} \right). \tag{A.0.13}$$

Here c is a universal constant.

Lemma A.5 (Bernstein-type Inequality, Proposition 5.17 of [Ver12]) *Let $X_1, \cdots, X_N$ be independent centered sub-exponetial random variables, and let $K = \max_i \left\|X_i\right\|_{\psi_1}$, where the sub-exponential norm*

$$\left\|X_i\right\|_{\psi_1} \doteq \sup_{p\geq 1} p^{-1} \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}. \tag{A.0.14}$$

Then for every $\boldsymbol{b} = [b_1; \cdots; b_N] \in \mathbb{C}^N$ and every $t \geq 0$, we have

$$\mathbb{P}\left(\left| \sum_{k=1}^{N} b_k X_k \right| \geq t \right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{K^2 \left\|\boldsymbol{b}\right\|_2^2}, \frac{t}{K \left\|\boldsymbol{b}\right\|_\infty} \right) \right). \tag{A.0.15}$$

Here c is a universal constant.

Lemma A.6 (Subgaussian Lower Tail for Nonnegative RV's, Problem 2.9 of [BLM13]) *Let $X_1, \ldots, X_N$ be i.i.d. copies of the nonnegative random variable X with finite second moment. Then it holds that*

$$\mathbb{P}\left[\frac{1}{N} \sum_{i=1}^{N} \left(X_i - \mathbb{E}\left[X_i\right]\right) < -t \right] \leq \exp\left(-\frac{Nt^2}{2\sigma^2} \right)$$

for any $t > 0$, where $\sigma^2 = \mathbb{E}\left[X^2\right]$.

Proof For any $\lambda > 0$, we have

$$\log \mathbb{E}\left[\mathrm{e}^{-\lambda(X - \mathbb{E}[X])} \right] = \lambda \mathbb{E}\left[X\right] + \log \mathbb{E}\left[\mathrm{e}^{-\lambda X}\right] \leq \lambda \mathbb{E}\left[X\right] + \mathbb{E}\left[\mathrm{e}^{-\lambda X}\right] - 1,$$

where the last inequality holds thanks to $\log u \leq u - 1$ for all $u > 0$. Moreover, using the fact $\mathrm{e}^u \leq 1 + u + u^2/2$ for all $u \leq 0$, we obtain

$$\log \mathbb{E}\left[\mathrm{e}^{-\lambda(X - \mathbb{E}[X])}\right] \leq \frac{1}{2}\lambda^2 \mathbb{E}\left[X^2\right] \Longleftrightarrow \mathbb{E}\left[\mathrm{e}^{-\lambda(X - \mathbb{E}[X])}\right] \leq \exp\left(\frac{1}{2}\lambda^2 \mathbb{E}\left[X^2\right]\right).$$

Thus, by the usual exponential transform trick, we obtain that for any $t > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{N}(X_i - \mathbb{E}\left[X_i\right]) < -t\right] \leq \exp\left(-\lambda t + N\lambda^2 \mathbb{E}\left[X^2\right]/2\right).$$

Taking $\lambda = t/(N\sigma^2)$ and making change of variable for t give the claimed result. ■

Appendix B

Auxillary Results for Sparse Dictionary Learning

In this chapter, we record supporting calculations and technical results for proofs of Part II.

B.1 Technical tools and basic facts

Lemma B.1 (Derivates and Lipschitz Properties of $h_\mu(z)$) *For the sparsity surrogate*

$$h_\mu(z) = \mu \log\left(\cosh\left(z/\mu\right)\right),$$

the first two derivatives are

$$\dot{h}_\mu(z) = \tanh\left(\frac{z}{\mu}\right), \quad \ddot{h}_\mu(z) = \frac{1}{\mu}\left[1 - \tanh^2\left(\frac{z}{\mu}\right)\right]. \tag{B.1.1}$$

Also, for any $z > 0$, we have

$$\frac{1}{2}\left(1 - \exp\left(-\frac{2z}{\mu}\right)\right) \;\le\; \tanh\left(\frac{z}{\mu}\right) \;\le\; 1 - \exp\left(-\frac{2z}{\mu}\right), \tag{B.1.2}$$

$$\exp\left(-\frac{2z}{\mu}\right) \;\le\; 1 - \tanh^2\left(\frac{z}{\mu}\right) \;\le\; 4\exp\left(-\frac{2z}{\mu}\right). \tag{B.1.3}$$

Moreover, for any $z,\ z' \in \mathbb{R}$, we have

$$\left|\dot{h}_\mu(z) - \dot{h}_\mu(z')\right| \le \frac{1}{\mu}\left|z - z'\right|, \quad \left|\ddot{h}_\mu(z) - \ddot{h}_\mu(z')\right| \le \frac{2}{\mu^2}\left|z - z'\right| \tag{B.1.4}$$

Lemma B.2 (Chebyshev's Association Inequality) *Let X denote a real-valued random variable, and $f, g : \mathbb{R} \mapsto \mathbb{R}$ nondecreasing (nonincreasing) functions of X with $\mathbb{E}[f(X)] < \infty$ and $\mathbb{E}[g(X)] < \infty$. Then*

$$\mathbb{E}[f(X)\, g(X)] \geq \mathbb{E}[f(X)]\, \mathbb{E}[g(X)]\,. \tag{B.1.5}$$

If f is nondecreasing (nonincreasing) and g is nonincreasing (nondecreasing), we have

$$\mathbb{E}[f(X)\, g(X)] \leq \mathbb{E}[f(X)]\, \mathbb{E}[g(X)]\,. \tag{B.1.6}$$

Proof Consider Y, an independent copy of X. Then it is easy to see

$$\mathbb{E}[(f(X) - f(Y))(g(X) - g(Y))] \geq 0.$$

Expanding the expectation and noticing $\mathbb{E}[f(X)\, g(Y)] = \mathbb{E}[f(Y)\, g(X)] = \mathbb{E}[f(X)]\, \mathbb{E}[g(X)]$ and also $\mathbb{E}[f(X)\, g(X)] = \mathbb{E}[f(Y)\, g(Y)]$ yields the result. Similarly, we can prove the second one. ■

This lemma implies the following lemma.

Lemma B.3 (Harris' Inequality, [Har60], see also Theorem 2.15 of [BLM13]) *Let $X_1, \ldots, X_n$ be independent, real-valued random variables and $f, g : \mathbb{R}^n \mapsto \mathbb{R}$ be nonincreasing (nondecreasing) w.r.t. any one variable while fixing the others. Define a random vector $\boldsymbol{X} = (X_1, \cdots, X_n) \in \mathbb{R}^n$, then we have*

$$\mathbb{E}[f(\boldsymbol{X})\, g(\boldsymbol{X})] \geq \mathbb{E}[f(\boldsymbol{X})]\, \mathbb{E}[g(\boldsymbol{X})]\,. \tag{B.1.7}$$

Similarly, if f is nondecreasing (nonincreasing) and g is nonincreasing (nondecreasing) coordinatewise in the above sense, we have

$$\mathbb{E}[f(\boldsymbol{X})\, g(\boldsymbol{X})] \leq \mathbb{E}[f(\boldsymbol{X})]\, \mathbb{E}[g(\boldsymbol{X})]\,. \tag{B.1.8}$$

Proof Again, it suffices to prove the first equality, which can be shown by induction. For $n = 1$, it reduces to Lemma B.2. Suppose the claim is true for any $m < n$. Since both g and f are nondecreasing functions in X_n given $\widehat{\boldsymbol{X}} = (X_1, \cdots, X_{n-1})$, then

$$\mathbb{E}[f(\boldsymbol{X})\, g(\boldsymbol{X})] = \mathbb{E}\left[\mathbb{E}\left[f(\boldsymbol{X})g(\boldsymbol{X}) \mid \widehat{\boldsymbol{X}}\right]\right] \geq \mathbb{E}\left[\mathbb{E}\left[f(\boldsymbol{X}) \mid \widehat{\boldsymbol{X}}\right] \mathbb{E}\left[g(\boldsymbol{X}) \mid \widehat{\boldsymbol{X}}\right]\right]$$

Now, it follows by independence that $f'\left(\widehat{\boldsymbol{X}}\right) = \mathbb{E}\left[f(\boldsymbol{X}) \mid \widehat{\boldsymbol{X}}\right]$ and $g'\left(\widehat{\boldsymbol{X}}\right) = \mathbb{E}\left[g(\boldsymbol{X}) \mid \widehat{\boldsymbol{X}}\right]$ are both nondecreasing functions, then by the induction hypothesis, we have

$$\mathbb{E}\left[f\left(\boldsymbol{X}\right)g\left(\boldsymbol{X}\right)\right] \geq \mathbb{E}\left[f'\left(\widehat{\boldsymbol{X}}\right)\right]\mathbb{E}\left[g'\left(\widehat{\boldsymbol{X}}\right)\right] = \mathbb{E}\left[f(\boldsymbol{X})\right]\mathbb{E}\left[g(\boldsymbol{X})\right],$$

as desired. ■

Lemma B.4 (Differentiation under the Integral Sign) *Consider a function* $F : \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}$ *such that* $\frac{\partial F(\boldsymbol{x},s)}{\partial s}$ *is well defined and measurable over* $\mathcal{U} \times (0, t_0)$ *for some open subset* $\mathcal{U} \subset \mathbb{R}^n$ *and some* $t_0 > 0$. *For any probability measure* μ *on* $\mathbb{R}^n$ *and any* $t \in (0, t_0)$ *such that* $\int_0^t \int_{\mathcal{U}} \left|\frac{\partial F(\boldsymbol{x},s)}{\partial s}\right| \mu\left(d\boldsymbol{x}\right) ds < \infty$, *it holds that*

$$\frac{d}{dt}\int_{\mathcal{U}} F\left(\boldsymbol{x},t\right)\mu\left(d\boldsymbol{x}\right) = \int_{\mathcal{U}} \frac{\partial F\left(\boldsymbol{x},t\right)}{\partial t}\mu\left(d\boldsymbol{x}\right), \text{ or } \frac{d}{dt}\mathbb{E}_{\boldsymbol{x}}\left[F\left(\boldsymbol{x},t\right)\mathbb{1}_{\mathcal{U}}\right] = \mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial F\left(\boldsymbol{x},t\right)}{\partial t}\mathbb{1}_{\mathcal{U}}\right]. \qquad \text{(B.1.9)}$$

Proof We have

$$\begin{aligned}\int_{\mathcal{U}} \frac{\partial F\left(\boldsymbol{x},t\right)}{\partial t}\mu\left(d\boldsymbol{x}\right) &= \frac{d}{dt}\int_0^t \int_{\mathcal{U}} \frac{\partial F\left(\boldsymbol{x},s\right)}{\partial s}\mu\left(d\boldsymbol{x}\right)ds \\ &= \frac{d}{dt}\int_{\mathcal{U}}\int_0^t \frac{\partial F\left(\boldsymbol{x},s\right)}{\partial s}\, ds\, \mu\left(d\boldsymbol{x}\right) \\ &= \frac{d}{dt}\int_{\mathcal{U}} \left(F\left(\boldsymbol{x},t\right) - F\left(\boldsymbol{x},0\right)\right)\, \mu\left(d\boldsymbol{x}\right) \\ &= \frac{d}{dt}\int_{\mathcal{U}} F\left(\boldsymbol{x},t\right)\, \mu\left(d\boldsymbol{x}\right),\end{aligned}$$

where we have used the fundamental theorem of calculus for the first and third equalities, and measure-theoretic Fubini's theorem (see, e.g., Theorem 2.37 of [Fol99]) for the second equality (as justified by our integrability assumption). ■

Lemma B.5 (Gaussian Tail Estimates) *Let* $X \sim \mathcal{N}(0,1)$ *and* $\Phi(x)$ *be CDF of* X. *For any* $x \geq 0$, *we have the following estimates for* $\Phi^c(x) \doteq 1 - \Phi(x)$:

$$\begin{aligned}\left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \leq \Phi^c\left(x\right) &\leq \left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5}\right)\frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}, \quad (\textit{Type I}) \qquad &\text{(B.1.10)} \\ \frac{x}{x^2+1}\frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \leq \Phi^c\left(x\right) &\leq \frac{1}{x}\frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}, \quad (\textit{Type II}) \qquad &\text{(B.1.11)} \\ \frac{\sqrt{x^2+4}-x}{2}\frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \leq \Phi^c\left(x\right) &\leq \left(\sqrt{2+x^2}-x\right)\frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \quad (\textit{Type III}). \qquad &\text{(B.1.12)}\end{aligned}$$

Proof Type I bounds can be obtained by integration by parts with proper truncations. Type II upper bound can again be obtained via integration by parts, and the lower bound can be obtained via considering the function $f(x) \doteq \Phi^c(x) - \frac{x}{x^2+1}\frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}$ and noticing it is always nonnegative. Type III bounds are mentioned in [Due10] and reproduced by the systematic approach developed therein (Section 2). ■

Lemma B.6 (Moments of the Gaussian Random Variables) *If $X \sim \mathcal{N}\left(0, \sigma^2\right)$, then it holds for all integer $p \geq 1$ that*

$$\mathbb{E}\left[|X|^p\right] = \sigma^p \left(p-1\right)!! \left[\sqrt{\frac{2}{\pi}} \mathbb{1}_{p \text{ odd}} + \mathbb{1}_{p \text{ even}}\right] \leq \sigma^p \left(p-1\right)!!. \tag{B.1.13}$$

Lemma B.7 (Moments of the χ^2 Random Variables) *If $X \sim \chi^2\left(n\right)$, then it holds for all integer $p \geq 1$,*

$$\mathbb{E}\left[X^p\right] = 2^p \frac{\Gamma\left(p+n/2\right)}{\Gamma\left(n/2\right)} = \prod_{k=1}^{p} (n+2k-2) \leq \frac{p!}{2}\left(2n\right)^p. \tag{B.1.14}$$

Lemma B.8 (Moments of the χ Random Variables) *If $X \sim \chi\left(n\right)$, then it holds for all integer $p \geq 1$,*

$$\mathbb{E}\left[X^p\right] = 2^{p/2} \frac{\Gamma\left(p/2+n/2\right)}{\Gamma\left(n/2\right)} \leq p! n^{p/2}. \tag{B.1.15}$$

Lemma B.9 (Integral Form of Taylor's Theorem) *Let $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ be a twice continuously differentiable function, then for any direction $\boldsymbol{y} \in \mathbb{R}^n$, we have*

$$f(\boldsymbol{x}+t\boldsymbol{y}) = f(\boldsymbol{x}) + t\int_0^1 \left\langle \nabla f(\boldsymbol{x}+st\boldsymbol{y}), \boldsymbol{y}\right\rangle \, ds, \tag{B.1.16}$$

$$f(\boldsymbol{x}+t\boldsymbol{y}) = f(\boldsymbol{x}) + t\left\langle \nabla f(\boldsymbol{x}), \boldsymbol{y}\right\rangle + t^2 \int_0^1 (1-s)\left\langle \nabla^2 f(\boldsymbol{x}+st\boldsymbol{y})\boldsymbol{y}, \boldsymbol{y}\right\rangle \, ds. \tag{B.1.17}$$

Proof By the fundamental theorem of calculus, since f is continuous differentiable, it is obvious that

$$f(\boldsymbol{x}+t\boldsymbol{y}) = f(\boldsymbol{x}) + \int_0^t \left\langle \nabla f(\boldsymbol{x}+\tau\boldsymbol{y}), \boldsymbol{y}\right\rangle d\tau. \tag{B.1.18}$$

If f is twice continuously differentiable, by using integral by parts, we obtain

$$\begin{aligned} f(\boldsymbol{x}+t\boldsymbol{y}) &= f(\boldsymbol{x}) + \left[(\tau-t)\left\langle \nabla f(\boldsymbol{x}+\tau\boldsymbol{y}), \boldsymbol{y}\right\rangle\right]\Big|_0^t - \int_0^t (\tau-t)\, d\left\langle \nabla f(\boldsymbol{x}+\tau\boldsymbol{y}), \boldsymbol{y}\right\rangle \\ &= f(\boldsymbol{x}) + t\left\langle \nabla f(\boldsymbol{x}+\tau\boldsymbol{y}), \boldsymbol{y}\right\rangle + \int_0^t (t-\tau)\left\langle \nabla^2 f(\boldsymbol{x}+\tau\boldsymbol{y})\boldsymbol{y}, \boldsymbol{y}\right\rangle d\tau. \end{aligned} \tag{B.1.19}$$

By a change of variable $\tau = st$ $(0 \leq s \leq 1)$ for (B.1.18) and (B.1.19), we get the desired results. ■

B.2 Auxiliary results

B.2.1 Integrals

Lemma B.10 *Let $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y \sim \mathcal{N}(0, \sigma_Y^2)$ be independent random variables and*

$$\Phi^c(t) \doteq \frac{1}{\sqrt{2\pi}} \int_t^\infty \exp\left(-x^2/2\right)\, dx$$

be the complementary cumulative distribution function of the standard normal. For any $a > 0$, we have

$$\mathbb{E}\left[X \mathbb{1}_{X>0}\right] = \frac{\sigma_X}{\sqrt{2\pi}}, \tag{B.2.1}$$

$$\mathbb{E}\left[\exp\left(-aX\right) X \mathbb{1}_{X>0}\right] = \frac{\sigma_X}{\sqrt{2\pi}} - a\sigma_X^2 \exp\left(\frac{a^2\sigma_X^2}{2}\right) \Phi^c\left(a\sigma_X\right), \tag{B.2.2}$$

$$\mathbb{E}\left[\exp\left(-aX\right) \mathbb{1}_{X>0}\right] = \exp\left(\frac{a^2\sigma_X^2}{2}\right) \Phi^c\left(a\sigma_X\right), \tag{B.2.3}$$

$$\begin{aligned}\mathbb{E}\left[\exp\left(-a(X+Y)\right) X^2 \mathbb{1}_{X+Y>0}\right] &= \sigma_X^2\left(1 + a^2\sigma_X^2\right) \exp\left(\frac{a^2\sigma_X^2 + a^2\sigma_Y^2}{2}\right) \Phi^c\left(a\sqrt{\sigma_X^2 + \sigma_Y^2}\right) \\ &\quad - \frac{a\sigma_X^4}{\sqrt{2\pi}\sqrt{\sigma_X^2 + \sigma_Y^2}},\end{aligned} \tag{B.2.4}$$

$$\begin{aligned}\mathbb{E}\left[\exp\left(-a(X+Y)\right) XY \mathbb{1}_{X+Y>0}\right] &= a^2\sigma_X^2\sigma_Y^2 \exp\left(\frac{a^2\sigma_X^2 + a^2\sigma_Y^2}{2}\right) \Phi^c\left(a\sqrt{\sigma_X^2 + \sigma_Y^2}\right) \\ &\quad - \frac{a\sigma_X^2\sigma_Y^2}{\sqrt{2\pi}\sqrt{\sigma_X^2 + \sigma_Y^2}},\end{aligned} \tag{B.2.5}$$

$$\mathbb{E}\left[\tanh\left(aX\right) X\right] = a\sigma_X^2 \mathbb{E}\left[1 - \tanh^2\left(aX\right)\right], \tag{B.2.6}$$

$$\mathbb{E}\left[\tanh\left(a(X+Y)\right) X\right] = a\sigma_X^2 \mathbb{E}\left[1 - \tanh^2\left(a(X+Y)\right)\right]. \tag{B.2.7}$$

Proof Equalities (B.2.1), (B.2.2), (B.2.3), (B.2.4) and (B.2.5) can be obtained by direct integrations. Equalities (B.2.6) and (B.2.7) can be derived using integration by part. ■

B.2.2 Proof of Lemma 9.1

Proof Indeed $\frac{1}{(1+\beta t)^2} = \sum_{k=0}^\infty (-1)^k (k+1) \beta^k t^k$, as

$$\sum_{k=0}^\infty (-1)^k (k+1) \beta^k t^k = \sum_{k=0}^\infty (-\beta t)^k + \sum_{k=0}^\infty k(-\beta t)^k = \frac{1}{1+\beta t} + \frac{-\beta t}{(1+\beta t)^2} = \frac{1}{(1+\beta t)^2}.$$

The magnitude of the coefficient vector is

$$\|\boldsymbol{b}\|_{\ell^1} = \sum_{k=0}^\infty \beta^k (1+k) = \sum_{k=0}^\infty \beta^k + \sum_{k=0}^\infty k\beta^k = \frac{1}{1-\beta} + \frac{\beta}{(1-\beta)^2} = \frac{1}{(1-\beta)^2} = T.$$

Observing that $\frac{1}{(1+\beta t)^2} > \frac{1}{(1+t)^2}$ for $t \in [0,1]$ when $0 < \beta < 1$, we obtain

$$\|p - f\|_{L^1[0,1]} = \int_0^1 |p(t) - f(t)|\, dt = \int_0^1 \left[\frac{1}{(1+\beta t)^2} - \frac{1}{(1+t)^2}\right] dt = \frac{1-\beta}{2(1+\beta)} \le \frac{1}{2\sqrt{T}}. \tag{B.2.8}$$

Moreover, we have

$$\|f-p\|_{L^\infty[0,1]} = \max_{t\in[0,1]} p(t)-f(t) = \max_{t\in[0,1]} \frac{t(1-\beta)\left(2+t(1+\beta)\right)}{(1+t)^2(1+\beta t)^2} \le 1-\beta = \frac{1}{\sqrt{T}}. \tag{B.2.9}$$

Finally, notice that

$$\begin{aligned}\sum_{k=0}^{\infty} \frac{b_k}{(1+k)^3} = \sum_{k=0}^{\infty} \frac{(-\beta)^k}{(1+k)^2} &= \sum_{i=0}^{\infty}\left[\frac{\beta^{2i}}{(1+2i)^2} - \frac{\beta^{2i+1}}{(2i+2)^2}\right] \\ &= \sum_{i=0}^{\infty} \beta^{2i} \frac{(2i+2)^2-\beta(2i+1)^2}{(2i+2)^2(2i+1)^2} > 0, \end{aligned} \tag{B.2.10}$$

where at the second equality we have grouped consecutive even-odd pair of summands. In addition, we have

$$\sum_{k=0}^{n} \frac{b_k}{(1+k)^3} \le \sum_{k=0}^{n} \frac{|b_k|}{(1+k)^3} = \sum_{k=0}^{n} \frac{\beta^k}{(1+k)^2} \le 1 + \sum_{k=1}^{n} \frac{1}{(1+k)k} = 2 - \frac{1}{n+1}, \tag{B.2.11}$$

which converges to 2 when $n \to \infty$, completing the proof. ■

B.2.3 Proof of of Lemma 9.4

Proof The first inequality is obviously true for $\boldsymbol{v} = \boldsymbol{0}$. When $\boldsymbol{v} \neq \boldsymbol{0}$, we have

$$\begin{aligned}\mathbb{E}\left[|\boldsymbol{v}^*\boldsymbol{z}|^m\right] &= \sum_{\ell=0}^{n} \theta^\ell (1-\theta)^{n-\ell} \sum_{\mathcal{J}\in\binom{[n]}{\ell}} \mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{v}_{\mathcal{J}}\|^2\right)}\left[|Z|^m\right] \\ &\le \sum_{\ell=0}^{n} \theta^\ell (1-\theta)^{n-\ell} \sum_{\mathcal{J}\in\binom{[n]}{\ell}} \mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{v}\|^2\right)}\left[|Z|^m\right] \\ &= \mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{v}\|^2\right)}\left[|Z|^m\right] \sum_{\ell=0}^{n} \theta^\ell (1-\theta)^{n-\ell} \binom{n}{\ell} \\ &= \mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{v}\|^2\right)}\left[|Z|^m\right], \end{aligned}$$

where the second line relies on the fact $\|\boldsymbol{v}_{\mathcal{J}}\| \le \|\boldsymbol{v}\|$ and that for a fixed order, central moment of Gaussian is monotonically increasing w.r.t. its variance. Similarly, to see the second inequality,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{z}\|^m\right] &= \sum_{\ell=0}^{n} \theta^\ell (1-\theta)^{n-\ell} \sum_{\mathcal{J}\in\binom{[n]}{\ell}} \mathbb{E}\left[\left\|\boldsymbol{z}'_{\mathcal{J}}\right\|^m\right] \\ &\le \mathbb{E}\left[\|\boldsymbol{z}'\|^m\right] \sum_{\ell=0}^{n} \theta^\ell (1-\theta)^{n-\ell} \binom{n}{\ell} = \mathbb{E}\left[\|\boldsymbol{z}'\|^m\right], \end{aligned}$$

as desired. ■

B.2.4 Proof of Lemma 9.11

Proof Consider one component of $\boldsymbol{X}$, i.e., $X_{ij} = B_{ij}V_{ij}$ for $i \in [n]$ and $j \in [p]$, where $B_{ij} \sim \mathrm{Ber}\,(\theta))$ and $V_{ij} \sim \mathcal{N}(0,1)$. We have

$$\mathbb{P}\left[|X_{ij}| > 4\sqrt{\log\,(np)}\right] \le \theta\mathbb{P}\left[|V_{ij}| > 4\sqrt{\log(np)}\right] \le \theta\exp\left(-8\log(np)\right) = \theta(np)^{-8}.$$

And also

$$\mathbb{P}\left[|X_{ij}| < 1\right] = 1 - \theta + \theta\mathbb{P}\left[|V_{ij}| < 1\right] \le 1 - 0.3\theta.$$

Applying a union bound as

$$\mathbb{P}\left[\|\boldsymbol{X}\|_\infty \le 1 \text{ or } \|\boldsymbol{X}\|_\infty \ge 4\sqrt{\log\,(np)}\right] \le (1-0.3\theta)^{np} + np\theta\,(np)^{-8} \le \exp\left(-0.3\theta np\right) + \theta\,(np)^{-7},$$

we complete the proof. ■

B.2.5 Matrix half-inverse perturbation bound

Lemma B.11 *Suppose $\boldsymbol{A} \succ \boldsymbol{0}$. Then for any symmetric perturbation matrix $\boldsymbol{\Delta}$ with $\|\boldsymbol{\Delta}\| \le \frac{\sigma_{\min}(\boldsymbol{A})}{2}$, it holds that*

$$\left\|(\boldsymbol{A}+\boldsymbol{\Delta})^{-1/2} - \boldsymbol{A}^{-1/2}\right\| \le \frac{2\left\|\boldsymbol{A}\right\|^{1/2}\|\boldsymbol{\Delta}\|}{\sigma^2_{\min}\left(\boldsymbol{A}\right)}. \tag{B.2.12}$$

Proof First note that

$$\left\|(\boldsymbol{A}+\boldsymbol{\Delta})^{-1/2} - \boldsymbol{A}^{-1/2}\right\| \le \frac{\left\|(\boldsymbol{A}+\boldsymbol{\Delta})^{-1} - \boldsymbol{A}^{-1}\right\|}{\sigma^{1/2}_{\min}\left(\boldsymbol{A}^{-1}\right)}$$

as by our assumption $\boldsymbol{A}+\boldsymbol{\Delta} \succ \boldsymbol{0}$ and the fact (Theorem 6.2 in [Hig08]) that

$$\left\|\boldsymbol{X}^{1/2} - \boldsymbol{Y}^{1/2}\right\| \le \|\boldsymbol{X}-\boldsymbol{Y}\| \,/\, \left(\sigma^{1/2}_{\min}\left(\boldsymbol{X}\right) + \sigma^{1/2}_{\min}\left(\boldsymbol{Y}\right)\right) \quad \text{for all} \quad \boldsymbol{X},\boldsymbol{Y} \succ \boldsymbol{0}$$

applies. Moreover, using the fact

$$\left\|(\boldsymbol{X}+\boldsymbol{\Delta})^{-1} - \boldsymbol{X}^{-1}\right\| \le \frac{\left\|\boldsymbol{X}^{-1}\right\|\left\|\boldsymbol{X}^{-1}\boldsymbol{\Delta}\right\|}{1-\left\|\boldsymbol{X}^{-1}\boldsymbol{\Delta}\right\|} \le \frac{\|\boldsymbol{\Delta}\|\left\|\boldsymbol{X}^{-1}\right\|^2}{1-\left\|\boldsymbol{X}^{-1}\right\|\|\boldsymbol{\Delta}\|}$$

for nonsingular $\boldsymbol{X}$ and perturbation $\boldsymbol{\Delta}$ with $\left\|\boldsymbol{X}^{-1}\right\|\|\boldsymbol{\Delta}\| < 1$ (see, e.g., Theorem 2.5 of Chapter III in [SS90]), we obtain

$$\tfrac{1}{\sigma_{\min}^{1/2}(\boldsymbol{A}^{-1})}\left\|(\boldsymbol{A}+\boldsymbol{\Delta})^{-1}-\boldsymbol{A}^{-1}\right\| \le \left\|\boldsymbol{A}\right\|^{1/2}\frac{\left\|\boldsymbol{\Delta}\right\|\left\|\boldsymbol{A}^{-1}\right\|^{2}}{1-\left\|\boldsymbol{A}^{-1}\right\|\left\|\boldsymbol{\Delta}\right\|} \le \frac{2\left\|\boldsymbol{A}\right\|^{1/2}\left\|\boldsymbol{\Delta}\right\|}{\sigma_{\min}^{2}\left(\boldsymbol{A}\right)},$$

where we have used the fact $\left\|\boldsymbol{A}^{-1}\right\|\left\|\boldsymbol{\Delta}\right\| \le 1/2$ to simplify at the last inequality. ■

B.2.6 BG matrix spectral estimate

> **Lemma B.12** *There exists a positive constant C such that for any $\theta \in (0, 1/2)$ and $n_2 > Cn_1^2 \log n_1$, the random matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ with $\boldsymbol{X} \sim_{i.i.d.} \mathrm{BG}\left(\theta\right)$ obeys*
>
> $$\left\|\frac{1}{n_2\theta}\boldsymbol{X}\boldsymbol{X}^* - \boldsymbol{I}\right\| \le 10\sqrt{\frac{\theta n_1 \log n_2}{n_2}} \tag{B.2.13}$$
>
> *with probability at least $1 - n_2^{-8}$.*

Proof Observe that $\mathbb{E}\left[\frac{1}{\theta}\boldsymbol{x}_k\boldsymbol{x}_k^*\right] = \boldsymbol{I}$ for any column $\boldsymbol{x}_k$ of $\boldsymbol{X}$ and so $\frac{1}{n_2\theta}\boldsymbol{X}\boldsymbol{X}^*$ can be considered as a normalize sum of independent random matrices. Moreover, for any integer $m \ge 2$,

$$\mathbb{E}\left[\left(\frac{1}{\theta}\boldsymbol{x}_k\boldsymbol{x}_k^*\right)^m\right] = \frac{1}{\theta^m}\mathbb{E}\left[\left\|\boldsymbol{x}_k\right\|^{2m-2}\boldsymbol{x}_k\boldsymbol{x}_k^*\right].$$

Now $\mathbb{E}\left[\left\|\boldsymbol{x}_k\right\|^{2m-2}\boldsymbol{x}_k\boldsymbol{x}_k^*\right]$ is a diagonal matrix (as $\mathbb{E}\left[\left\|\boldsymbol{x}_k\right\|^2 x_k\left(i\right)x_k\left(j\right)\right] = -\mathbb{E}\left[\left\|\boldsymbol{x}_k\right\|^2 x_k\left(i\right)x_k\left(j\right)\right]$ for any $i \ne j$ by symmetry of the distribution) in the form $\mathbb{E}\left[\left\|\boldsymbol{x}_k\right\|^{2m-2}\boldsymbol{x}_k\boldsymbol{x}_k^*\right] = \mathbb{E}\left[\left\|\boldsymbol{x}\right\|^{2m-2}x(1)^2\right]\boldsymbol{I}$ for $\boldsymbol{x} \sim_{i.i.d.} \mathrm{BG}\left(\theta\right)$ with $\boldsymbol{x} \in \mathbb{R}^{n_1}$. Let $t^2\left(\boldsymbol{x}\right) = \left\|\boldsymbol{x}\right\|^2 - x(1)^2$. Then if $m = 2$,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}\right\|^2 x(1)^2\right] &= \mathbb{E}\left[x(1)^4\right] + \mathbb{E}\left[t^2\left(\boldsymbol{x}\right)\right]\mathbb{E}\left[x(1)^2\right] \\ &= \mathbb{E}\left[x(1)^4\right] + \left(n_1 - 1\right)\left(\mathbb{E}\left[x(1)^2\right]\right)^2 = 3\theta + \left(n_1 - 1\right)\theta^2 \le 3n_1\theta,\end{aligned}$$

where for the last simplification we use the assumption $\theta \le 1/2$. For $m \ge 3$,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}\right\|^{2m-2}x(1)^2\right] &= \sum_{k=0}^{m-1}\binom{m-1}{k}\mathbb{E}\left[t^{2k}\left(\boldsymbol{x}\right)x(1)^{2m-2k}\right] = \sum_{k=0}^{m-1}\binom{m-1}{k}\mathbb{E}\left[t^{2k}\left(\boldsymbol{x}\right)\right]\mathbb{E}\left[x(1)^{2m-2k}\right] \\ &\le \sum_{k=0}^{m-1}\binom{m-1}{k}\mathbb{E}_{Z\sim\chi^2(n_1-1)}\left[Z^k\right]\theta\mathbb{E}_{W\sim\mathcal{N}(0,1)}\left[W^{2m-2k}\right] \\ &\le \theta\sum_{k=0}^{m-1}\binom{m-1}{k}\frac{k!}{2}\left(2n_1-2\right)^k\left(2m-2k\right)!!\end{aligned}$$

$$\begin{aligned}
&\le \theta 2^m \frac{m!}{2} \sum_{k=0}^{m-1} \binom{m-1}{k} (n_1 - 1)^k \\
&\le \frac{m!}{2} n_1^{m-1} 2^{m-1},
\end{aligned}$$

where we have used the moment estimates for Gaussian and χ^2 random variables from Lemma B.6 and Lemma B.7, and also $\theta \le 1/2$. Taking $\sigma^2 = 3n_1\theta$ and $R = 2n_1$, and invoking the matrix Bernstein in Lemma A.2, we obtain

$$\mathbb{E}\left[\left\| \frac{1}{p\theta} \sum_{k=1}^{p} \boldsymbol{x}_k \boldsymbol{x}_k^* - \boldsymbol{I} \right\| > t\right] \le \exp\left(- \frac{n_2 t^2}{6 n_1 \theta + 4 n_1 t} + 2 \log n_1 \right) \tag{B.2.14}$$

for any $t \ge 0$. Taking $t = 10\sqrt{\theta n_1 \log(n_2)/n_2}$ gives the claimed result. ■

B.2.7 Subspace angles and distance

> **Lemma B.13** *Consider two linear subspaces $\mathcal{U}$, $\mathcal{V}$ of dimension k in $\mathbb{R}^n$ ($k \in [n]$) spanned by orthonormal bases $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Suppose $\pi/2 \ge \theta_1 \ge \theta_2 \cdots \ge \theta_k \ge 0$ are the principal angles between $\mathcal{U}$ and $\mathcal{V}$. Then it holds that*
>
> *i)* $\min_{\boldsymbol{Q} \in O_k} \|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{Q}\| \le \sqrt{2 - 2\cos\theta_1}$;
>
> *ii)* $\sin\theta_1 = \|\boldsymbol{U}\boldsymbol{U}^* - \boldsymbol{V}\boldsymbol{V}^*\|$;
>
> *iii) Let $\mathcal{U}^\perp$ and $\mathcal{V}^\perp$ be the orthogonal complement of $\mathcal{U}$ and $\mathcal{V}$, respectively. Then $\theta_1(\mathcal{U}, \mathcal{V}) = \theta_1(\mathcal{U}^\perp, \mathcal{V}^\perp)$.*

Proof Proof to i) is similar to that of II. Theorem 4.11 in [SS90]. For $2k \le n$, w.l.o.g., we can assume $\boldsymbol{U}$ and $\boldsymbol{V}$ are the canonical bases for $\mathcal{U}$ and $\mathcal{V}$, respectively. Then

$$\min_{\boldsymbol{Q} \in O_k} \left\| \begin{bmatrix} \boldsymbol{I} - \boldsymbol{\Gamma}\boldsymbol{Q} \\ -\boldsymbol{\Sigma}\boldsymbol{Q} \\ \boldsymbol{0} \end{bmatrix} \right\| \le \left\| \begin{bmatrix} \boldsymbol{I} - \boldsymbol{\Gamma} \\ -\boldsymbol{\Sigma} \\ \boldsymbol{0} \end{bmatrix} \right\| \le \left\| \begin{bmatrix} \boldsymbol{I} - \boldsymbol{\Gamma} \\ -\boldsymbol{\Sigma} \end{bmatrix} \right\|.$$

Now by definition

$$\begin{aligned}
\left\| \begin{bmatrix} \boldsymbol{I} - \boldsymbol{\Gamma} \\ -\boldsymbol{\Sigma} \end{bmatrix} \right\|^2 &= \max_{\|\boldsymbol{x}\|=1} \left\| \begin{bmatrix} \boldsymbol{I} - \boldsymbol{\Gamma} \\ -\boldsymbol{\Sigma} \end{bmatrix} \boldsymbol{x} \right\|^2 = \max_{\|\boldsymbol{x}\|=1} \sum_{i=1}^{k} (1 - \cos\theta_i)^2 x_i^2 + \sin^2\theta_i x_i^2 \\
&= \max_{\|\boldsymbol{x}\|=1} \sum_{i=1}^{k} (2 - 2\cos\theta_i) x_i^2 \le 2 - 2\cos\theta_1.
\end{aligned}$$

Note that the upper bound is achieved by taking $\boldsymbol{x} = \boldsymbol{e}_1$. When $2k > n$, by the results from CS decomposition (see, e.g., I Theorem 5.2 of [SS90]).

$$\min_{\boldsymbol{Q}\in O_k} \left\| \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Gamma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} \boldsymbol{I} - \boldsymbol{\Gamma} \\ -\boldsymbol{\Sigma} \end{bmatrix} \right\|,$$

and the same argument then carries through. To prove ii), note the fact that $\sin\theta_1 = \|\boldsymbol{U}\boldsymbol{U}^* - \boldsymbol{V}\boldsymbol{V}^*\|$ (see, e.g., Theorem 4.5 and Corollary 4.6 of [SS90]). Obviously one also has

$$\sin\theta_1 = \|\boldsymbol{U}\boldsymbol{U}^* - \boldsymbol{V}\boldsymbol{V}^*\| = \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^*) - (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^*)\|,$$

while $\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^*$ and $\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^*$ are projectors onto $\mathcal{U}^\perp$ and $\mathcal{V}^\perp$, respectively. This completes the proof. ■

Appendix C

Auxillary Results for Generalized Phase Retrieval

In this chapter, we record supporting calculations and technical results for proofs of Part III.

> **Lemma C.1 (Even Moments of Complex Gaussian)** *For $a \sim \mathcal{CN}(1)$, it holds that*
>
> $$\mathbb{E}\left[|a|^{2p}\right] = p! \quad \forall\, p \in \mathbb{N}.$$

Proof Write $a = x + \mathrm{i}y$, then $x, y \sim_{i.i.d.} \mathcal{N}(0, 1/2)$. Thus,

$$\mathbb{E}\left[|a|^{2p}\right] = \mathbb{E}_{x,y}\left[\left(x^2+y^2\right)^p\right] = \frac{1}{2^p}\mathbb{E}_{z\sim\chi^2(2)}\left[z^p\right] = \frac{1}{2^p}2^p p! = p!,$$

as claimed. ■

> **Lemma C.2 (Integral Form of Taylor's Theorem)** *Consider any continuous function $f(\boldsymbol{z}) : \mathbb{C}^n \mapsto \mathbb{R}$ with continuous first- and second-order Wirtinger derivatives. For any $\boldsymbol{\delta} \in \mathbb{C}^n$ and scalar $t \in \mathbb{R}$, we have*
>
> $$f(\boldsymbol{z}+t\boldsymbol{\delta}) = f(\boldsymbol{z}) + t\int_0^1 \begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix}^* \nabla f(\boldsymbol{z}+st\boldsymbol{\delta})\, ds,$$
>
> $$f(\boldsymbol{z}+t\boldsymbol{\delta}) = f(\boldsymbol{z}) + t\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix}^* \nabla f(\boldsymbol{z}) + t^2\int_0^1 (1-s)\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix}^* \nabla^2 f(\boldsymbol{z}+st\boldsymbol{\delta})\begin{bmatrix}\boldsymbol{\delta}\\ \overline{\boldsymbol{\delta}}\end{bmatrix}\, ds.$$

Proof Since f is continuous differentiable, by the fundamental theorem of calculus,

$$f(\boldsymbol{z} + t\boldsymbol{\delta}) = f(\boldsymbol{z}) + \int_0^t \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z} + \tau\boldsymbol{\delta}) \; d\tau.$$

Moreover, by integral by part, we obtain

$$\begin{aligned} f(\boldsymbol{z} + t\boldsymbol{\delta}) &= f(\boldsymbol{z}) + \left[(\tau - t) \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z} + \tau\boldsymbol{\delta}) \right] \Bigg|_0^t - \int_0^t (\tau - t) \; d \left[\begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z} + \tau\boldsymbol{\delta}) \right] \\ &= f(\boldsymbol{x}) + t \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z}) + \int_0^t (t - \tau) \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z} + \tau\boldsymbol{\delta}) \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix} d\tau. \end{aligned}$$

Change of variable $\tau = st(0 \le s \le 1)$ gives the claimed result. ■

> **Lemma C.3 (Error of Quadratic Approximation)** *Consider any continuous function $f(\boldsymbol{z}) : \mathbb{C}^n \mapsto \mathbb{R}$ with continuous first- and second-order Wirtinger derivatives. Suppose its Hessian $\nabla^2 f(\boldsymbol{z})$ is L_h-Lipschitz. Then the second-order approximation*
>
> $$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{z}) = f(\boldsymbol{z}) + \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla f(\boldsymbol{z}) + \frac{1}{2} \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}$$
>
> *around each point $\boldsymbol{z}$ obeys*
>
> $$\left| f(\boldsymbol{z} + \boldsymbol{\delta}) - \widehat{f}(\boldsymbol{\delta}; \boldsymbol{z}) \right| \le \frac{1}{3} L_h \left\| \boldsymbol{\delta} \right\|^3.$$

Proof By integral form of Taylor's theorem in Lemma C.2,

$$\begin{aligned} \left| f(\boldsymbol{z} + \boldsymbol{\delta}) - \widehat{f}(\boldsymbol{\delta}; \boldsymbol{z}) \right| &= \left| \int_0^1 (1 - \tau) \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix}^* \left[\nabla^2 f(\boldsymbol{x} + \tau\boldsymbol{\delta}) - \nabla^2 f(\boldsymbol{x}) \right] \begin{bmatrix} \boldsymbol{\delta} \\ \overline{\boldsymbol{\delta}} \end{bmatrix} \; d\tau \right| \\ &\le \; 2 \left\| \boldsymbol{\delta} \right\|^2 \int_0^1 (1 - \tau) \left\| \nabla^2 f(\boldsymbol{x} + \tau\boldsymbol{\delta}) - \nabla^2 f(\boldsymbol{x}) \right\| \; d\tau \\ &\le \; 2 L_h \left\| \boldsymbol{\delta} \right\|^3 \int_0^1 (1 - \tau)\tau \; d\tau = \frac{L_h}{3} \left\| \boldsymbol{\delta} \right\|^3, \end{aligned}$$

as desired. ■

Lemma C.4 (Spectrum of Complex Gaussian Matrices) *Let $\boldsymbol{X}$ be an $n_1 \times n_2$ ($n_1 > n_2$) matrices with i.i.d. $\mathcal{CN}$ entries. Then,*

$$\sqrt{n_1} - \sqrt{n_2} \leq \mathbb{E}\left[\sigma_{\min}(\boldsymbol{X})\right] \leq \mathbb{E}\left[\sigma_{\max}(\boldsymbol{X})\right] \leq \sqrt{n_1} + \sqrt{n_2}.$$

Moreover, for each $t \geq 0$, it holds with probability at least $1 - 2\exp\left(-t^2\right)$ that

$$\sqrt{n_1} - \sqrt{n_2} - t \leq \sigma_{\min}(\boldsymbol{X}) \leq \sigma_{\max}(\boldsymbol{X}) \leq \sqrt{n_1} + \sqrt{n_2} + t.$$